

**STRUCTURE-BASED METHODS FOR THE PHYLOGENETIC ANALYSIS OF
RIBOSOMAL RNA MOLECULES**

A Dissertation

by

JOSEPH JAMES GILLESPIE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Entomology

**STRUCTURE-BASED METHODS FOR THE PHYLOGENETIC ANALYSIS OF
RIBOSOMAL RNA MOLECULES**

A Dissertation

by

JOSEPH JAMES GILLESPIE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,
Committee Members,

Head of Department,

Anthony Cognato
James Woolley
Keyan Zhu-Salzman
Rodney Honeycutt
Kevin Heinz

August 2005

Major Subject: Entomology

ABSTRACT

Structure-Based Methods for the Phylogenetic Analysis of Ribosomal RNA Molecules.

(August 2005)

Joseph James Gillespie, B.S, Widener University;

M.S., University of Delaware

Chair of Advisory Committee: Dr. Anthony Cognato

Ribosomal RNA (rRNA) molecules form highly conserved secondary and tertiary structures via rRNA-rRNA and rRNA-protein interactions that collectively comprise the macromolecule that is the ribosome. Because of their cellular universality, rRNA molecules are commonly used for phylogeny estimations spanning all divergences of life. In this dissertation, I elucidate the structure of several rRNAs by analyzing multiply aligned sequences for basepair covariation and conserved higher order structural motifs. Specifically, I predict novel structures for expansion segments D2 and D3 of the nuclear large subunit rRNA (28S) and variable regions V4-V9 of the nuclear small subunit rRNA (18S) from 249 galerucine leaf beetles (Coleoptera: Chrysomelidae). I describe a novel means for characterizing regions of alignment ambiguity that improves methods for retaining phylogenetic information without violating nucleotide positional homology. In the program PHASE, I explore a variety of RNA maximum likelihood models using the 28S rRNA dataset and discuss the utility of these models in light of their performance under Bayesian analysis. I conclude that seven-state models are likely the best models to use for phylogenetic estimation, although I cannot determine with

confidence which of the two seven-state models (7A or 7D) is better. Evaluation of the unpaired sites within both rRNAs in Modeltest provided a similar model of evolution for these non-pairing regions (TrN+ I+G). In addition, a sequenced region of the mitochondrial cytochrome oxidase I gene (COI) from the galerucines was evaluated in Modeltest, with each codon position modeled separately (GTR+I+G for positions 1 and 2, GTR+G for position 3). The combined galerucine dataset (28S+18S rRNA helices, 28S+18S rRNA unpaired sites, COI 1st, 2nd and 3rd positions) provided for two mixed-model Bayesian analysis of five discretely-modeled partitions (using 7A and 7D). The results of these analyses are compared with those obtained from equally weighted parsimony to provide a robust phylogenetic estimate of the Galerucinae and related leaf beetle taxa. Finally, the odd characteristics of strepsipteran 18S rRNA are evaluated through comparison of 12 strepsipterans with 163 structurally-aligned arthropod sequences. Among other interesting results, I identify errors in previously published strepsipteran sequences and predict structures not previously known from metazoan rRNA.

To my dear wife Annika,
thanks for your patience and constant
support of my work on rRNA

In loving memory of the National Hockey League and mullets (and the Ramones)

This entire dissertation is dedicated to Explosions In The Sky (the earth is not a cold dead place), tastykakes (a Philly thing), our garden, every TOOL album, jelly bellies, rain, Michael Moore, “2+2=5”, Miller High Life (mostly tall boys), David Chappelle, the decaf (?) coffee at the Ag Cafe, my U2 !pod, Frehley, Gwen, the fishies (dead and alive), Antonio's Pizza, Jeff Buckley, the cigarettes I used to smoke, Donnie Darko (movie and soundtrack), Microsoft Word (kidding), MHs ~ 5:17, The Arcade Fire, Ralph Nader, the Binaural album (especially “Insignificance”), and, of course, ribosomal RNA.

ACKNOWLEDGMENTS

I thank my committee members, Anthony Cognato, Rodney Honeycutt, Jim Woolley and Keyan Zhu-Salzman, for invaluable mentoring and assistance with many aspects of my research, and their selflessness and patience.

I also thank my family and friends for their love and continued support of my research endeavors.

I acknowledge the following collaborators of mine: Jamie Cannone, Shawn Clarke, Anthony Cognato, Andy Deans, Rakesh Dhanda, Catherine Duckett, Brian Farrell, Robin Gutell, David Hawks, John Heraty, Rodney Honeycutt, Spencer Johnston, Jeya Kathirithamby, Karl Kjer, Hans Klompen, Hernan Lopez, David Maddison, John Mallat, Jody Martin, Claire McKenna, James Munro, Karen Ober, Ed Riley, Doug Tallamy, Erik J. van Nieukerken, Bob Wharton, Brian Wiegmann, Don Windsor, Jim Woolley, Matt Yoder.

Finally, I express my gratitude to the Department of Entomology, Texas A&M University, for providing me with the assistance and invaluable resources to complete this dissertation.

TABLE OF CONTENTS

	Page
ABSTRACT	iii
DEDICATION	v
ACKNOWLEDGMENTS.....	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES.....	ix
LIST OF TABLES	xiii
 CHAPTER	
I INTRODUCTION	1
II A SECONDARY STRUCTURAL MODEL OF THE 28S rRNA EXPANSION SEGMENTS D2 AND D3 FROM ROOTWORMS AND RELATED LEAF BEETLES (COLEOPTERA: CHRYSOMELIDAE; GALERUCINAE).....	5
Overview	5
Introduction	6
Results and discussion.....	9
Experimental procedures.....	40
III CHARACTERIZING REGIONS OF AMBIGUOUS ALIGNMENT CAUSED BY THE EXPANSION AND CONTRACTION OF HAIRPIN-STEM LOOPS IN RIBOSOMAL RNA MOLECULES.....	50
Introduction	50
Results and discussion.....	59
Experimental procedures.....	66
IV INCORPORATING MAXIMUM LIKELIHOOD MODELS FOR HELICAL AND NON-PAIRING REGIONS OF RIBOSOMAL RNA IN PHYLOGENETIC ANALYSIS: AN EXAMPLE FROM THE 28S LSU rRNA D2 AND D3 EXPANSION SEGMENTS OF	

CHAPTER	Page
ROOTWORMS AND RELATED LEAF BEETLES (COLEOPTERA: CHRYSOMELIDAE; GALERUCINAE).....	68
Overview	68
Introduction	70
Results and discussion.....	81
Conclusion.....	112
Experimental procedures.....	114
V PHYLOGENY OF ROOTWORMS AND RELATED GALERUCINE BEETLES (COLEOPTERA: CHRYSOMELIDAE) BASED ON THE ANALYSIS OF PARTIAL 28S AND 18S rRNA AND COI GENE SEQUENCES.....	118
Overview	118
Introduction	119
Results and discussion.....	132
Experimental procedures.....	166
VI ASSESSING THE ODD STRUCTURAL PROPERTIES OF NUCLEAR SMALL SUBUNIT RIBOSOMAL RNA SEQUENCES (18S) OF THE TWISTED-WING PARASITES (INSECTA: STREPSIPTERA)	189
Overview	189
Introduction	190
Results and discussion.....	197
Experimental procedures.....	222
VII SUMMARY	226
REFERENCES.....	228
VITA	253

LIST OF FIGURES

FIGURE	Page
1 A schematic line drawing of the secondary structure of LSU 28S rRNA from the beetle <i>Tenebrio</i> sp.	11
2 Multiple sequence alignment of primary and secondary structure of the expansion segments D2 and D3 of the LSU 28S nuclear rRNA gene from six chrysomelid species	12
3 The secondary structure model of the expansion segments D2 and D3 of the LSU 28S nuclear rRNA gene from spotted cucumber beetle (<i>Diabrotica undecimpunctata howardi</i>)	15
4 A gallery of diverse secondary structure diagrams from the "helix 2" compound helix in the D2 region (synonymous with the 545 gallery of Schnare <i>et al.</i> (1996)) is shown for the following chrysomelid taxa: A. <i>Acalymma vittata</i> , B. <i>Agelastica coerulea</i> , C. <i>Cerochroa brachialis</i> , D. <i>Coptocycla adamantina</i> , E. <i>Epitrix fasciata</i> , F. <i>Lamprosoma</i> sp., G. <i>Metaxyonycha panamensis</i> , H. <i>Neolochmaea dilatipennis</i> , I. <i>Pyrrhalta aenescens</i> J. Thailand specimen 11, K. <i>Walterianella bucki</i>	17
5 A gallery of diverse secondary structure diagrams from the "helix 3-1" compound helix in the D2 region (synonymous with the 545 gallery of Schnare <i>et al.</i> (1996)) is shown for the following chrysomelid taxa: A. <i>Acalymma vittata</i> , B. <i>Agelastica coerulea</i> , C. <i>Cerochroa brachialis</i> , D. <i>Coptocycla adamantina</i> , E. <i>Epitrix fasciata</i> , F. <i>Lamprosoma</i> sp., G. <i>Metaxyonycha panamensis</i> , H. <i>Neolochmaea dilatipennis</i> , I. <i>Pyrrhalta aenescens</i> J. Thailand specimen 11, K. <i>Walterianella bucki</i> , L. <i>Eucerotoma</i> sp. 344.	18
6 A gallery of diverse secondary structure diagrams from the "helix 3-2" compound helix in the D2 region (synonymous with the 545 gallery of Schnare <i>et al.</i> (1996)) is shown for the following chrysomelid taxa: A. <i>Agelastica coerulea</i> , B. <i>Acalymma vittata</i> , C. <i>Cerochroa brachialis</i> , D. <i>Coptocycla adamantina</i> , E. <i>Epitrix fasciata</i> , F. <i>Lamprosoma</i> sp., G. <i>Metaxyonycha panamensis</i> , H. <i>Neolochmaea dilatipennis</i> , I. <i>Pyrrhalta aenescens</i> , J. Thailand specimen 11, K. <i>Walterianella bucki</i>	20

FIGURE

Page

7	A gallery of diverse secondary structure diagrams for the D3 region (synonymous with the 650 gallery of Schnare <i>et al.</i> (1996)) is shown for the following chrysomelid taxa: A. <i>Cerochroa brachialis</i> , B. <i>Scelidopsis</i> sp., C. <i>Coptocyclus adamantina</i> , D. <i>Epitrix fasciata</i> , E. <i>Lamprosoma</i> sp., F. <i>Metaxyonycha panamensis</i> , G. <i>Neolochmaea dilatipennis</i> , H. <i>Pyrrhalta aenescens</i> , I. Thailand specimen 11, J. <i>Mimastra gracilicornis</i>	21
8	Predicted secondary structure for the expansion segment D2 of the 28S rRNA from <i>Diabrotica undecimpunctata howardi</i> (Coleoptera: Chrysomelidae), GenBank accession number AY243738	58
9	Demonstration of the subdivision of ambiguously-aligned regions of rDNA sequences formed by the expansion and contraction of hairpin-stem loops	61
10	Definition of the rate matrix for the most general time reversible 16-state model of RNA evolution	74
11	Predicted secondary structure model of the expansion segments D2 and D3 of the large subunit ribosomal RNA 28S of 231 sampled chrysomelid leaf beetles.....	80
12	Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 6A	86
13	Plots of the four poor parameters over three and six generations for the analyses performed under model 6A	87
14	Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 6B	89
15	Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 7A	93
16	Plots of the ten poor parameters over three and six generations for the analyses performed under model 7A.....	94

FIGURE	Page
17 Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 7D	98
18 Superimposed branches of the trees generated under model 7A for three and six million sampling generations.....	110
19 Superimposed branches of the trees generated under model 7D for three and six million sampling generations.....	111
20 Consensus secondary structure diagram of domains II and III of the nuclear small subunit rRNA (18S) gene region for the chrysomelids sequenced here	141
21 Results of an equally-weighted parsimony analysis of the combined data (COI nucleotides, 18S and 28S rRNA nucleotides).....	143
22 Extended majority rule consensus from a Bayesian analysis of the combined data (COI nucleotides, 18S and 28S rRNA nucleotides, 18S and 28S rRNA basepairs) under five maximum likelihood models (one per each partition). The rRNA basepairs were modeled under model 7A	149
23 Extended majority rule consensus from a Bayesian analysis of the combined data (COI nucleotides, 18S and 28S rRNA nucleotides, 18S and 28S rRNA basepairs) under five maximum likelihood models (one per each partition). The rRNA basepairs were modeled under model 7D	156
24 Plot of the log likelihood ($\ln L$) over the sampled generations for the three and five million generation analyses performed on the combined data with rRNA basepairs modeled under model 7A	161
25 Plot of the log likelihood ($\ln L$) over the sampled generations for the three and five million generation analyses performed on the combined data with rRNA basepairs modeled under model 7D	162
26 The secondary structure model of the nuclear SSU rRNA (18S) from the strepsipteran <i>Caenocholax fenyesi texensis</i> (accession number DQ026302)	195

FIGURE		Page
27	A gallery of diverse secondary structure diagrams of the variable region 2 (V2) and related core elements from selected strepsipterans	204
28	A comparison between our predicted structures for helix H184c and those of Choe <i>et al.</i> (1999b).....	205
29	A gallery of diverse secondary structure diagrams of the insertion within the hairpin loop of pseudoknot 13/14 within variable region 4 (V4) from selected strepsipterans	208
30	Recreated MALIGN alignment of Whiting <i>et al.</i> (1997, pg. 56)	217

LIST OF TABLES

TABLE	Page
1 Composition and degree of compensation for the base pairs of the D2 and D3 expansion segments and related core regions of the 28S rRNA in rootworms and related chrysomelid beetles	24
2 Mean percent nucleotides and mean transition/transversion ratios in pairing (stems) and non-pairing (loops) regions of the D2 and D3 expansion segments of the 28S LSU gene of chrysomelids.....	31
3 A list of the 18 regions of alignment ambiguity (RAA), one region of slipped-strand compensation (RSC) and two regions of expansion and contraction (REC) created in the multiple sequence alignment of the expansion segments D2 and D3 of the 28S LSU rRNA from 229 sampled chrysomelids	32
4 Secondary structure characters of the D2, D3 expansion segments from the higher-level chrysomelid taxa sampled in this analysis	36
5 A list of the chrysomeloid taxa analyzed in this investigation.....	41
6 Glossary of terms used for alignment and secondary structure of rRNA	52
7 Proposed RNA maximum likelihood models (modified from Savill <i>et al.</i> , 2001).....	75
8 Maximum likelihood values and AIC values for the best models and mean scores of 3000 and 6000 samples from the posterior probability distribution.....	82
9 Mean and estimated samples sizes for model 6A statistics.....	84
10 Mean and estimated samples sizes for model 6B statistics.....	88
11 Mean and estimated samples sizes for model 7A statistics.....	91
12 Mean and estimated samples sizes for model 7D statistics.....	97

TABLE		Page
13	Best rate matrices of the models evaluated in this study.....	100
14	Rank of substitution types and frequencies estimated for six-state RNA models for basepairs within the helices of the 28S rRNA expansion segments and related core elements	102
15	Rank of substitution types and frequencies estimated for seven-state RNA models for basepairs within the helices of the 28S rRNA expansion segments and related core elements	103
16	Likelihood ratio tests performed within the six and seven-state models.....	107
17	Observed and model basepair frequencies within the helices of the 28S rRNA expansion segments and related core elements.....	108
18	Relative conservation of amino acid residues from the sampled region of the cytochrome oxidase I gene (COI)	133
19	Monophyletic groups recovered in the parsimony and Bayesian (PHASE) analyses	165
20	A list of the 249 chrysomeloid taxa analyzed in this investigation.....	167
21	A list of the oligonucleotide primers used to amplify and sequence the chrysomeloids analyzed in this investigation.....	177
22	Results from running Modeltest on the molecular partitions.....	181
23	Summary statistics on the variable regions of the 18S rRNA in the major arthropod classes and hexapod orders.....	198
24	Distribution of sequence and structure within the strepsipteran insertion in the second hairpin loop of pseudoknot 2 within the variable region 4 (V4) of the 18S rRNA molecule	212
25	Composition and degree of compensation for the base pairs of putative helices H184-2 and HV2-V4 of the 18S rRNA across 175 arthropods.....	214

CHAPTER I

INTRODUCTION

Ribosomal RNA molecules form secondary and tertiary structures that are highly conserved across divergent organisms, a consequence of the need for preservation of ribosomal function in cellular protein synthesis (Dahlberg, 1989; Wool *et al.*, 1990; Noller, 1991). Higher-order structure in rRNA is obtained primarily through base-pair interactions within the individual RNA molecule (Fresco *et al.*, 1960). Hydrogen-bonding occurs between canonical base-pairs (AU, GC), non-canonical stable (GU) and unstable (AC) intermediates, as well as uncommon GA and AA pairings (Elgavish *et al.*, 2001), to form contiguous, antiparallel structural elements (helices). Other less-frequently occurring basepairs, as well as other secondary structural elements and tertiary interactions are reviewed in Gutell *et al.* (1994; 2002), and, together with conserved secondary structural helices, form the universally-conserved core ribosome that is comparable across all domains of life (Woese *et al.*, 1990a; Winker & Woese, 1991). This organismal universality of the ribosome, coupled with other characteristics such as high copy number of rDNA cistrons per cell and relative ease for primer design in conserved RNA regions, account for the commonality of rDNA sequences as markers for phylogeny reconstruction across virtually any lineage of life (Hillis & Dixon, 1991).

This dissertation follows the style of *Insect Molecular Biology*.

The multiple sequence alignment of rDNA is often problematic when the degree of length heterogeneity amongst taxa is high (De Rijk *et al.*, 1995). While the majority of helices in rRNA molecules are structurally conserved across the most divergent of taxa (Gutell *et al.*, 1994; Gutell, 1996), some helices and non-pairing regions, such as hairpin-stem loops and terminal and lateral bulges, can vary greatly in nucleotide sequence length and base composition even in closely related taxa (e.g., Hillis & Dixon, 1991; Schnare *et al.*, 1996; Gillespie *et al.*, 2004b). This characteristic of rRNA structure, coupled with the fact that pairing and non-pairing regions often accumulate substitutions at different rates (Van de Peer *et al.*, 1993), suggests that evolutionary studies utilizing these molecules for phylogeny reconstruction should benefit from the *a priori* designation of higher order structure to rDNA sequences. For instance, several studies have shown that structural information provides an objective criterion for assigning positional nucleotide homology in difficult-to-align rDNA datasets (e.g., Kjer, 1995; Hickson *et al.*, 1996; Kjer, 1997; Noterdame *et al.*, 1997; Hwang *et al.*, 1998; Lutzoni *et al.*, 2000; Goertzen *et al.*, 2003; Xia *et al.*, 2003). Also, Hickson *et al.* (2000) demonstrated that automated alignment methods fail to align sequences according to their conserved structural motifs, undoubtedly a consequence of these algorithms being based on phenetic sequence distance as opposed to structures that are more conserved than primary nucleotide sequence. Some studies have even shown that alignments based on structural information improve phylogeny estimation (Dixon & Hillis, 1993; Kjer, 1995; Titus & Frost, 1996; Morrison & Ellis, 1997; Uchida *et al.*, 1998; Mugridge *et al.*, 1999; Cunningham *et al.*, 2000; Gonzalez & Labarere, 2000; Hwang & Kim, 2000;

Lydeard *et al.*, 2000; Morin, 2000; Xia, 2000; Xia *et al.*, 2003). A recent example of this is the study of Xia *et al.* (2003) in which only structural alignments (and appropriate substitution models based on structure) were able to recover the well-accepted phylogeny of tetrapods using "analytically-challenging" nuclear SSU rDNA (18S) sequences.

I assert here that phylogenetic studies using rRNA gene regions as markers should be performed in unison with higher order structure prediction of these molecules. These nucleotide sequences are not letters; they are nucleotides that contain inter- and intra-molecular basepairs that have been experimentally proven with not only covariation analysis, but also recent crystalline structures of the ribosome. Still, it is important for the reader of the chapters within this dissertation to realize that homology assignment in any set of sequences is purely hypothetical. I argue that biological criteria, such as covariation analysis (with subsequent statistical assessment), thermodynamic algorithms, and comparative evidence, are all objective means to improve the assignment of positional nucleotide homology, especially in sequence alignments that contain a high level of length heterogeneity.

In this dissertation, I predict novel structures for the expansion segments D2 and D3 of the nuclear large subunit rRNA (28S) and variable regions V4-V9 of the nuclear small subunit rRNA (18S) from 249 galerucine leaf beetles (Coleoptera: Chrysomelidae). I describe a novel means for characterizing regions of ambiguously-aligned sequences that improves methods for retaining phylogenetic information without violating nucleotide positional homology. In the program PHASE, I explore a variety of

RNA maximum likelihood models using the 28S rRNA dataset and discuss the utility of these models in light of their performance under Bayesian analysis. A combined galerucine dataset (28S+18S rRNA helices, 28S+18S rRNA unpaired sites, COI 1st, 2nd and 3rd positions) is analyzed under parsimony and two mixed-model Bayesian analyses of five discretely-modeled partitions (using 7A and 7D). These three analyses are then discussed regarding the phylogeny of the Galerucinae and related leaf beetle taxa. Finally, the odd characteristics of strepsipteran 18S rRNA are evaluated through the comparison of 12 strepsipterans with 163 structurally-aligned arthropod sequences. Among other interesting results, I identify errors in previously published strepsipteran sequences and elucidate predicted structures not previously known from metazoan rRNA.

I demonstrate here that structure can be predicted from multiple sequence alignments to: 1. improve homology assignment and provide an objective criterion for data exclusion (a "conditional combination" approach for phylogeny estimation), 2. provide information about the sequenced molecules that allows for sub-partitions of the datasets to be created and modeled as independent character classes ("stems and loops"), 3. improve the existing knowledge of the structure and function of rRNA and ultimately the ribosome, while often identifying novel structural features, and 4. identify sequencing artifacts on public genetic databases that were previously undetected without structural inference. My dissertation will be useful for evolutionary biologists concerned with the structure, function, and evolution of rRNA, as well as systematists interested in structure-based applications for the phylogenetic analysis of these intriguing molecules.

CHAPTER II

A SECONDARY STRUCTURAL MODEL OF THE 28S rRNA EXPANSION SEGMENTS D2 AND D3 FROM ROOTWORMS AND RELATED LEAF BEETLES (COLEOPTERA: CHRYSOMELIDAE; GALERUCINAE)*

Overview

We analyze the secondary structure of two expansion segments (D2, D3) of the 28S rRNA gene from 229 leaf beetles (Coleoptera: Chrysomelidae), the majority of which are in the subfamily Galerucinae. The sequences are compared in a multiple sequence alignment, with secondary structure inferred primarily from the compensatory base changes in the conserved helices of the rRNA molecules. This comparative approach yielded 30 helices that are comprised of base pairs with positional covariation. Based on these leaf beetle sequences, we report an annotated secondary structural model for the D2, D3 expansion segments that will prove useful in assigning positional nucleotide homology for phylogeny reconstruction in these and closely related beetle taxa. This predicted structure, consisting of seven major compound helices, is mostly consistent with previously proposed models for the D2 and D3 expansion segments in insects.

* This article, Gillespie, J.J., Cannone, J.J., Gutell, R.R., Cognato, A.I. (2004) A secondary structural model of the 28S rRNA expansion segments D2 and D3 from rootworms and related leaf beetles (Coleoptera: Chrysomelidae; Galerucinae). *Insect Mol Biol* **13**: 495-518, is reprinted with permission from Blackwell Publishing, copyright 2004.

Despite a lack of conservation in the primary structure of these regions of insect 28S rRNA, the evolution of the secondary structure of these seven major motifs may be informative above the nucleotide level for higher-order phylogeny reconstruction of major insect lineages.

Introduction

The nuclear-encoded ribosomal large subunit (LSU) rRNA-encoding gene (23S-like rRNA) varies greatly in sequence length and nucleotide composition within the main eukaryote lineages (Ware *et al.*, 1983; Clark *et al.*, 1984; Hassouna *et al.*, 1984). The length heterogeneity in eukaryotic lineages is isolated to specific regions of the LSU rRNA (Clark, 1987; Gorski, *et al.* 1987; Michot & Bachellerie, 1987; Hancock & Dover, 1988; Tautz *et al.*, 1988; Gutell & Fox, 1988), of which some are referred to as expansion segments (Clark *et al.*, 1984). While these regions of the rRNA are usually not associated with protein translation (Gerbi, 1985), site-directed mutagenesis studies have implicated one of these highly variable regions with function (Sweeney *et al.*, 1994). In addition, the structure in these regions with less sequence conservation and more length variation is more variable than the structure in the regions with more sequence conservation and less length variation.

The eukaryotic rDNA occurs as a multi-gene family of tandemly-repeated units of the 23S-like, 16S-like and 5.8S rRNA transcripts that evolve concertedly (Arnheim *et al.*, 1980; Dover, 1982; Arnheim, 1983; Flavell, 1986). These tandem arrays, called nucleolar organization regions (NORs), are located on chromosomes in hundreds to

thousands of copies throughout the genome, with copy number dependent on the organism in question. Unequal crossing over and gene conversion keep the many copies of NORs conserved within species (Dover, 1982). The three functional rRNA transcripts are separated by internally transcribed spacers (ITSs) that are spliced out of the transcripts after NOR expression. While all three transcripts contain regions of variability (in base composition and sequence length), the 23S-like transcript has 13 expansion segments, as well as nine other identified variable regions (Schnare *et al.*, 1996), of rapidly-evolving sequence and is the most variable of the nuclear rRNA genes (Mindell & Honeycutt, 1991). This variation is associated with a wide range of phylogenetically informative characters among higher taxonomic levels (De Rijk *et al.*, 1995; Schnare *et al.*, 1996; Kuzoff, *et al.* 1998).

The 13 expansion segments of the 28S rRNA vary greatly among insect orders (Hwang *et al.*, 1998; Gillespie, unpubl. data), as well as within Diptera (Tautz *et al.* 1988; Kjer *et al.*, 1994; Schnare *et al.*, 1996) and Hymenoptera (Belshaw & Quicke, 2002; Gillespie, unpubl. data). As in other eukaryotes, the expansion segments in insects are more variable than the core rRNA, but are constrained structurally, with deleterious mutations often accommodated by compensatory base changes that maintain helical formation (Hancock *et al.*, 1988; Tautz *et al.*, 1988; Rousset *et al.*, 1991; Kjer *et al.*, 1994). This duality of variability and conservation makes these regions ideal for phylogenetic reconstruction among insects because the variation yields phylogenetic information and structural conservation helps the assessment of nucleotide homology. For example, the 28S-D1 and D3 regions have been utilized in the reconstruction of

Trichoptera phylogeny (Kjer *et al.*, 2001), and the 28S-D2 region has been used to resolve tribal relationships within galerucine leaf beetles (Gillespie *et al.*, 2003, 2004). However, their use in phylogeny reconstruction of Insecta is often problematic due to the difficulty of alignment of multiple sequences from divergent taxa (De Rijk *et al.*, 1995). This problem derives from the variability within the expansion segments, particularly in the distal regions of expanding and contracting hairpin-stem loop motifs (Crease & Taylor, 1998; Gillespie, In press). Thus, unlike the alignment of highly conserved core regions of rRNA molecules, the expansion segments require inspection for compensatory base changes that facilitate the alignment of highly divergent sequences. Co-evolving helices and highly conserved single-stranded regions empirically provide homology assignments that delimit unalignable regions (Kjer, 1995; 1997). After initial exclusion, these subsequent alignment-ambiguous regions can be incorporated into phylogeny reconstruction in a variety of ways. They can be recoded as multistate characters based on nucleotide identity (Lutzoni *et al.*, 2000; Kjer *et al.*, 2001; Gillespie *et al.* 2003, 2004), and further subjected to a step matrix that implements unequivocal weighting to character transformations (Lutzoni *et al.*, 2000; Gillespie *et al.* 2003, 2004; Xia *et al.*, 2003; Sorenson *et al.*, 2003). Unalignable regions can also be recoded as morphological characters based on the differences these regions impose on the secondary structure of the molecule (Billoud *et al.*, 2000; Collins *et al.*, 2000; Lydeard *et al.*, 2000; Ouvard *et al.*, 2000). Across taxa, transformations from one structure to another can be calculated as a measure of structural variability (Fontana *et al.*, 1993; Notredame *et al.*, 1997; Moulton *et al.*, 2000; Misof & Fleck, 2003). Homologous, yet

unalignable structures can even be characterized as phylogenetic trees, with differences in tree topology representing transformations across variable structures (Shapiro & Zhang, 1990; Hofacker *et al.*, 1994).

In this study, we present a structural model for the expansion segments D2 and D3 of the 28S rRNA gene from 229 leaf beetles (Coleoptera: Chrysomelidae), the majority of which are found in the subfamily Galerucinae. This model is a refined annotation from previous studies that incorporated secondary structure to improve homology assignment for phylogeny reconstruction of these beetles (Gillespie, 2001; Gillespie *et al.*, 2003, 2004; Kim *et al.*, 2003). Using compensatory base change evidence, we define conserved regions of the molecule that provide a custom chrysomelid model for this region of the 28S rRNA gene. Our novel characterization of regions of alignment ambiguity (RAA), slipped-strand compensation (RSC) and expansion and contraction (REC) from structural homology is discussed within taxonomic and phylogenetic contexts. This model will be useful for future studies on related beetle groups that utilize the D2 and D3 expansion segments for phylogeny reconstruction, and for studies that address expansion segment evolution across higher-level insect taxa (Misof & Fleck, 2003).

Results and discussion

Predicted secondary structure

The first nearly complete predicted secondary structural model of the eukaryotic cytoplasmic LSU rRNA from a beetle, the tenebrionid *Tenebrio* sp., is shown here (Fig.

1) in concordance with the conserved 23S and 23S-like structures of the LSU rRNA from the literature (Wool, 1986; Gutell & Fox, 1988; Gutell *et al.*, 1990, 1992b, 1993; Schnare *et al.*, 1996). With existing predicted structures for *Drosophila melanogaster* (Schnare *et al.*, 1996, others therein), *Aedes albopictus* (Kjer *et al.*, 1994), and *Acyrtosiphon pisum* (Amako *et al.*, 1996), this is the fourth predicted structure of the 28S LSU rRNA from an insect. The expansion segments D2 and D3 are highlighted and correspond, respectively, to the variable regions 545 and 650 of Schnare *et al.* (1996), which refer to the sequence numbering of *E. coli* LSU rRNA (Fig. 1). A multiple sequence alignment spanning the two expansion segments was generated from 229 chrysomelid taxa; however, six sampled taxa are listed for brevity (Fig. 2). The entire alignment is posted in a variety of electronic formats at <http://hisl.tamu.edu>, <http://www.rna.icmb.utexas.edu/>, and on the *Insect Molecular Biology* website.

Of the 864 positions in the *D. undecimpunctata howardi* reference sequence, we have identified 676 nucleotide positions in the 28S-D2,D3 sequence alignment that can be confidently assigned positional homology across the beetle taxa. Of the remaining length-variable positions, 18 regions of alignment ambiguity (RAA), one region of slipped-strand compensation (RSC) and two regions of expansion and contraction (REC) were identified and excluded from primary homology assignment. The 30 conserved helices within the D2 and D3 expansion segments of the 28S rRNA gene are illustrated on a two-dimensional structural model, which also includes the core regions of the 28S between the D2 and D3 and flanking the D3 in the 3' direction (Fig. 3). Less compensatory base change evidence is found within the D3 expansion segment because

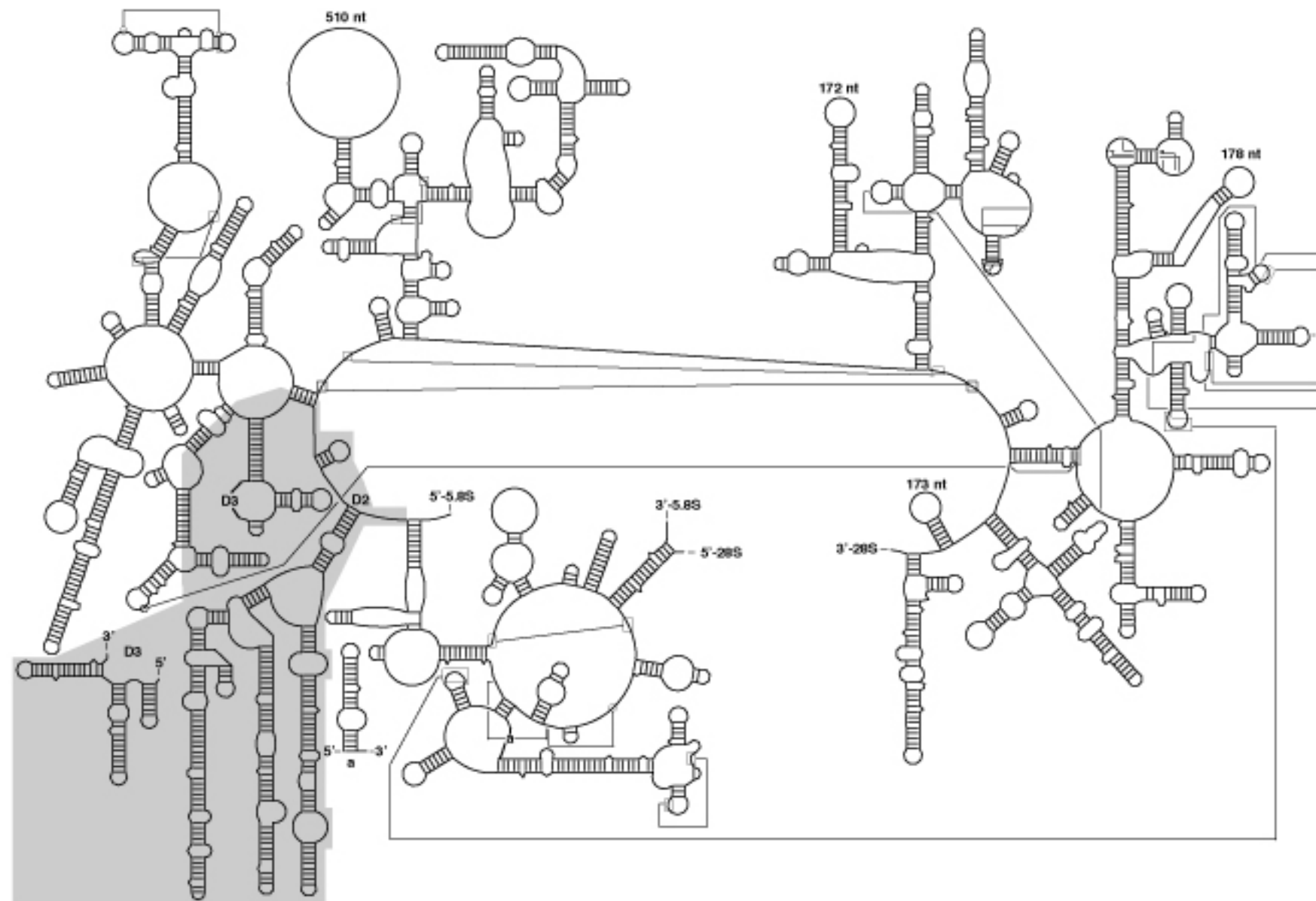


Figure 1. A schematic line drawing of the secondary structure of LSU 28S rRNA from the beetle *Tenebrio* sp. (accession number AY210843). The shaded region shows the expansion segments D2 and D3 (regions 545 and 650, respectively, of Schnare *et al.* (1996)) and related core sequence that were analyzed in this study. Base-pairing (where there is strong comparative support) and tertiary interactions that link the 5'- and 3'-halves of the molecule are shown connected by continuous lines. Structures for the expansion segments D7a, D7b, D8, D10, and D12 are preliminary at this time (most structures are shown as arcs or loops, with numbers indicating size). These structures will be adjusted when more beetle sequences from these regions are made available.

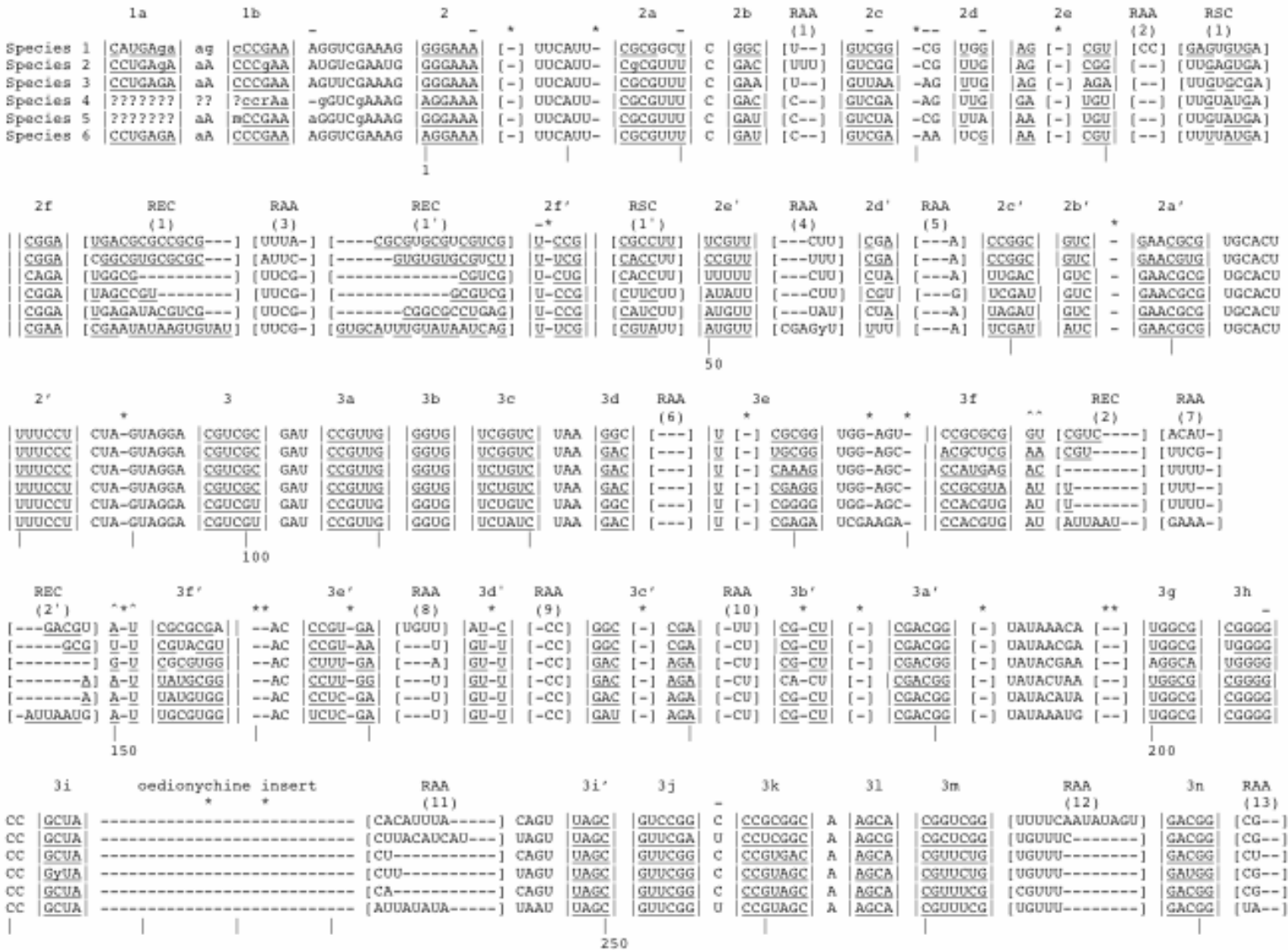


Figure 2. Multiple sequence alignment of primary and secondary structure of the expansion segments D2 and D3 of the LSU 28S nuclear rRNA gene from six chrysomelid species (*Lamprosoma* sp., *M. panamensis*, *E. fasciata*, *D. adelpha*, *P. aenescens*, *N. dilatipennis*). Regions of core rRNA between the two expansion segments and flanking the 3' end of the D3 are numbered following Cannone *et al.* (2002). The notation for the 26 conserved helices within the expansion segment D2 is modified from Gillespie *et al.* (2003a, b) with slight annotations to the previous predicted structure (See Figure 3). Helices with long range interactions are placed within bars (|) and immediate hairpin-stem loops are placed within double bars (||). All complimenatry strands are depicted with a prime ('); e.g., strand **1** hydrogen bonds with strand **1'** to form helix **1**). Regions of alignment ambiguity (**RAA**), slipped-strand compensation (**RSC**) and expansion and contraction (**REC**) are placed within brackets ([]). Nucleotides within helices involved in hydrogen-bonding are underlined. Single insertions (*) and deletions (-) are noted as in Kjer *et al.* (2001). Positions which can form an expansion of a helix across some but not all taxa are labeled with a caret (^). Every tenth nucleotide assigned positional homology is noted under the alignment with a tick (|), with every 50th position numbered. The sequences are 5' to 3' in direction. Missing nucleotides are represented with question marks (?). Lower case letters depict nucleotides confirmed by one strand only in sequencing. Note: this alignment has not been amended for these six taxa from the original alignment of 229 chrysomelid sequences, thus gaps and insertions may correspond to taxa not presented in this figure.

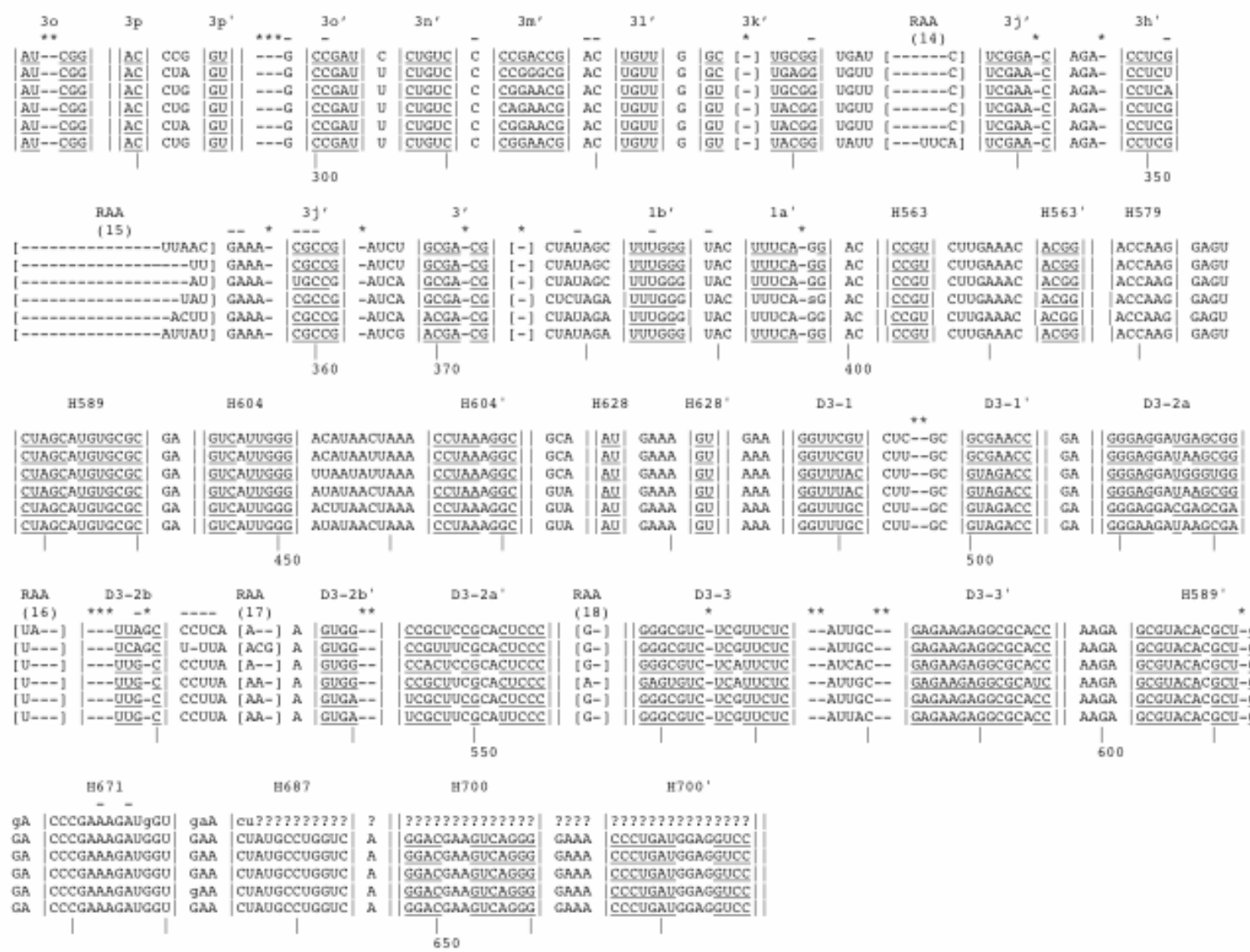


Figure 2 Continued.

many of the analyzed sequences are from studies that only included the D2 expansion segment (Gillespie *et al.*, 2003, 2004; Kim *et al.*, 2003).

Expansion segment D2

The 28S-D2 segment, corresponding to the 545 variable region of the 23S-like LSU (Schnare *et al.* 1996), is comprised of four main compound helices that are flanked by highly conserved elements in the 28S core structure. These motifs are labeled "helix 1", "helix 2", "helix 3-1" and "helix 3-2", while the sub-components of the compound helices are named a, b, c, etc. (Fig. 3). A total of 26 conserved helical elements comprise the D2 in chrysomelids (but see below regarding helix **3q** in *A. coerulea*). The innermost helix of the D2, named here as helices **1a** and **1b** (helix A in Schnare *et al.*, 1996), could not be evaluated for compensatory base changes due to the prevalence of unknown nucleotide assignments in electropherograms because of the close proximity of the 5'-primer to strand **1**.

Helix **2** in the D2 region is at the base of the second compound helix and is comprised of six basepairs across nearly all holometabolous insects (Gillespie, unpubl. data). The chrysomelids contain six helices that are apical to helix **2** (**2a-2f**). Many of the basepairs within these helices are supported with positional covariation. A gallery of structures representing the "helix 2" motif is presented in Figure 4. The terminal helix in this motif, helix **2f**, has the potential to form additional base-pairings beyond the four boxed basepairs; however, a confident homology assignment is not possible here due to the high sequence and length variation in this region (see REC 1 below). One RSC, one

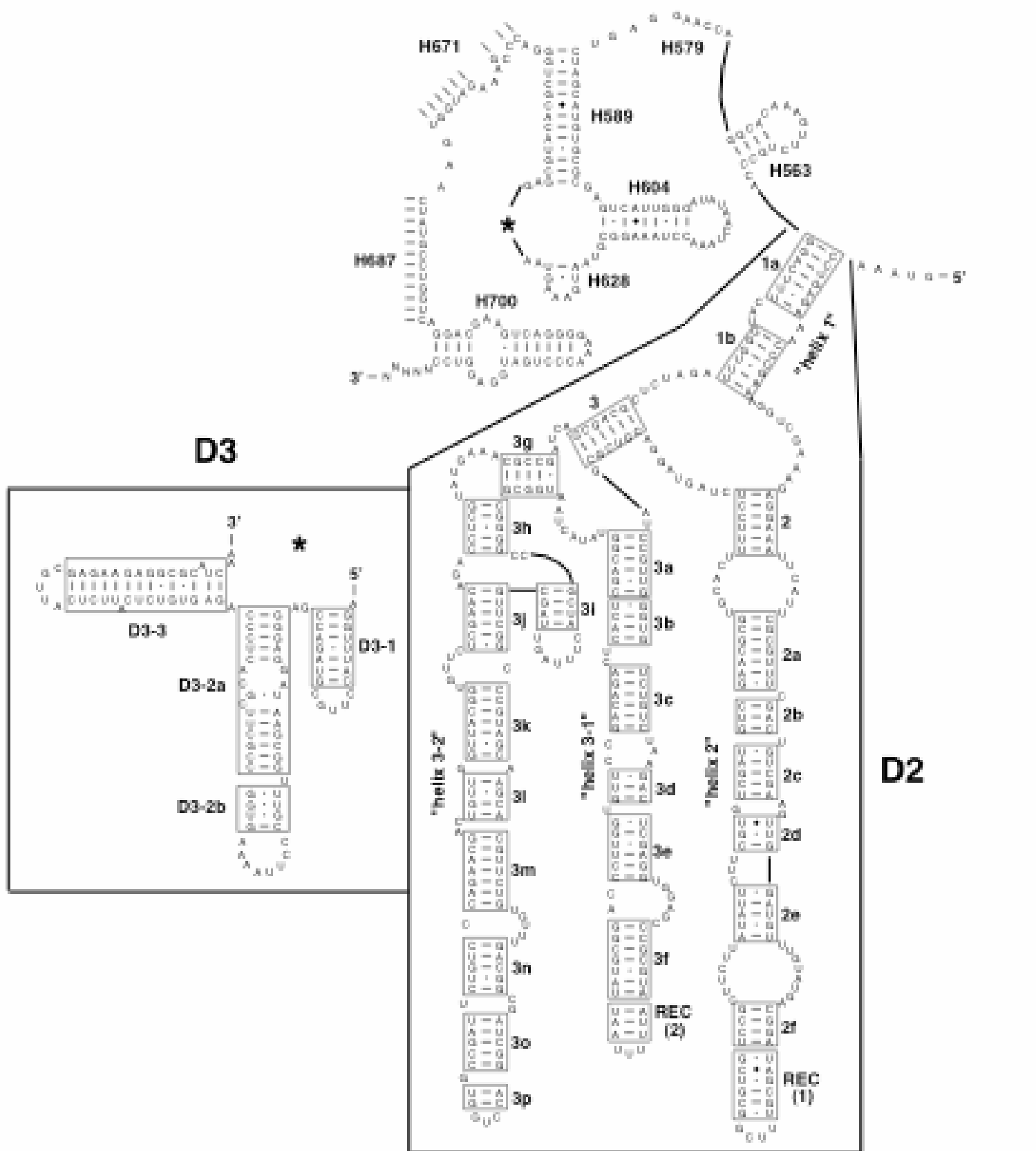


Figure 3. The secondary structure model of the expansion segments D2 and D3 of the LSU 28S nuclear rRNA gene from spotted cucumber beetle (*Diabrotica undecimpunctata howardi*). The 30 conserved, covarying helices present in all of the beetle taxa studied here are boxed. Helix notation is modified from Gillespie *et al.* (2003, 2004) (see Figure 2). Regions of core rRNA between the two expansion segments and flanking the 3' end of the D3 are numbered following Cannone *et al.* (2002). Base-pairing is indicated as follows: standard canonical pairs by lines (C-G, G-C, A-U, U-A); wobble G·U pairs by dots (G·U); A-G pairs by open circles (A^oG); other non-canonical pairs by filled circles (e.g. C•A). Diagram was generated using the program XRNA (Weiser, B. & Noller, H., University of California at Santa Cruz).

REC and six RAAs occur in "helix 2" (Fig. 4F).

Helix **3** (H2 in Michot & Bachellerie, 1997; E in Schnare *et al.*, 1996) is highly conserved in the higher eukaryotes and is the most basal helix to several compound helices (Schnare *et al.*, 1996; Gillespie, unpubl. data). Helix **3** is six basepairs long in the chrysomelids and most holometabolous insect lineages (Gillespie, unpubl. data). The chrysomelids have two compound helices distal to helix **3**, "helix 3-1" (helices **3a-3f**) and "helix 3-2" (helices **3g-3p**) (Fig. 3). A gallery of representative "helix 3-1" structures for different chrysomelids is displayed in Figure 5. The terminal helix in "helix 3-1", **3f**, has the potential to form additional base-pairings beyond the seven boxed positions; however, this homology assignment is ambiguous for the positions identified in REC (2) and RAA (7) (distal to the **3f** boxed basepairs in Fig. 5G) due to the lack of sequence conservation and the variation in sequence lengths. Although most taxa in the alignment append two more basepairs onto helix **3f**, the taxon *Eucerotoma* sp. 344 (Fig. 5L) has only seven basepairs in helix **3f**. Thus, we limited helix **3f** to seven basepairs because only these positions represent a homologous structure across the alignment. "Helix 3-1" has one REC and five RAAs (Fig. 5G).

A gallery of different chrysomelid "helix 3-2" compound helices are shown in Figure 6. Unlike the first two compound helices in the D2 expansion segment, which contain some length variation, the terminal helices of "helix 3-2", **3o** and **3p**, are very conserved in length and base composition. In contrast, helix **3i**, is variable in length (14-50 nts) and sequence across all taxa (e.g., Fig. 6K). Length variation is also located in the unpaired nucleotides between strands **3h'** and **3g'**, ranging from 4 to 24 nucleotides.

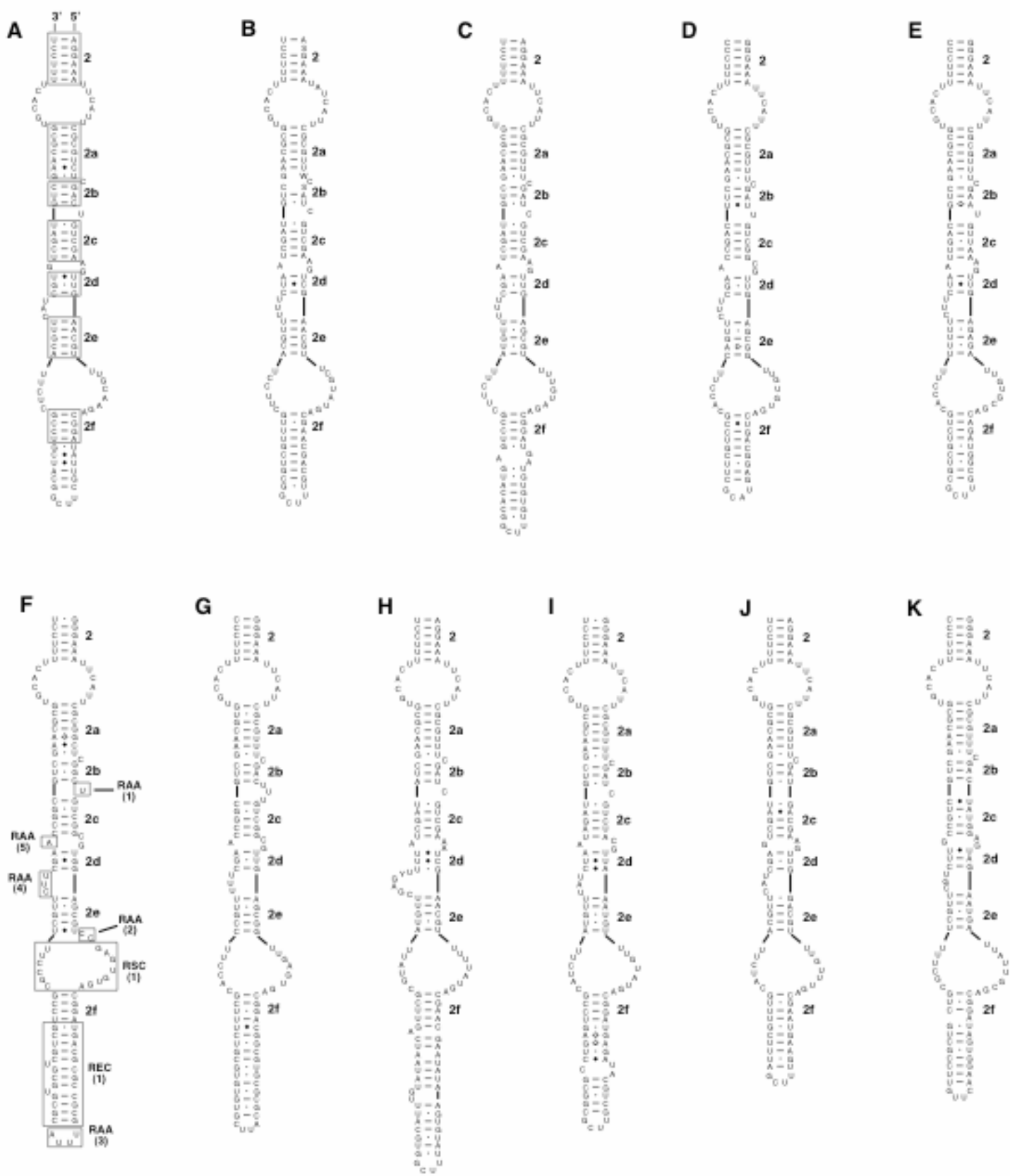


Figure 4. A gallery of diverse secondary structure diagrams from the "helix 2" compound helix in the D2 region (synonymous with the 545 gallery of Schnare *et al.* (1996)) is shown for the following chrysomelid taxa: A. *Acalymma vittata*, B. *Agelastica coerulea*, C. *Cerochroa brachialis*, D. *Coptocycla adamantina*, E. *Epitrix fasciata*, F. *Lamprosoma* sp., G. *Metaxyonycha panamensis*, H. *Neolochmaea dilatipennis*, I. *Pyrrhalta aenescens* J. Thailand specimen 11, K. *Walterianella bucki*. Notation for the seven helical elements is modified from Gillespie *et al.* (2003, 2004). Helices are boxed in A., and ambiguously-aligned regions are boxed in F. The notation for RAAs, RSCs and RECs is described in Figure 2 and on page 32. The explanations of base-pair symbols and reference for software used to construct structure diagrams are in Figure 3.

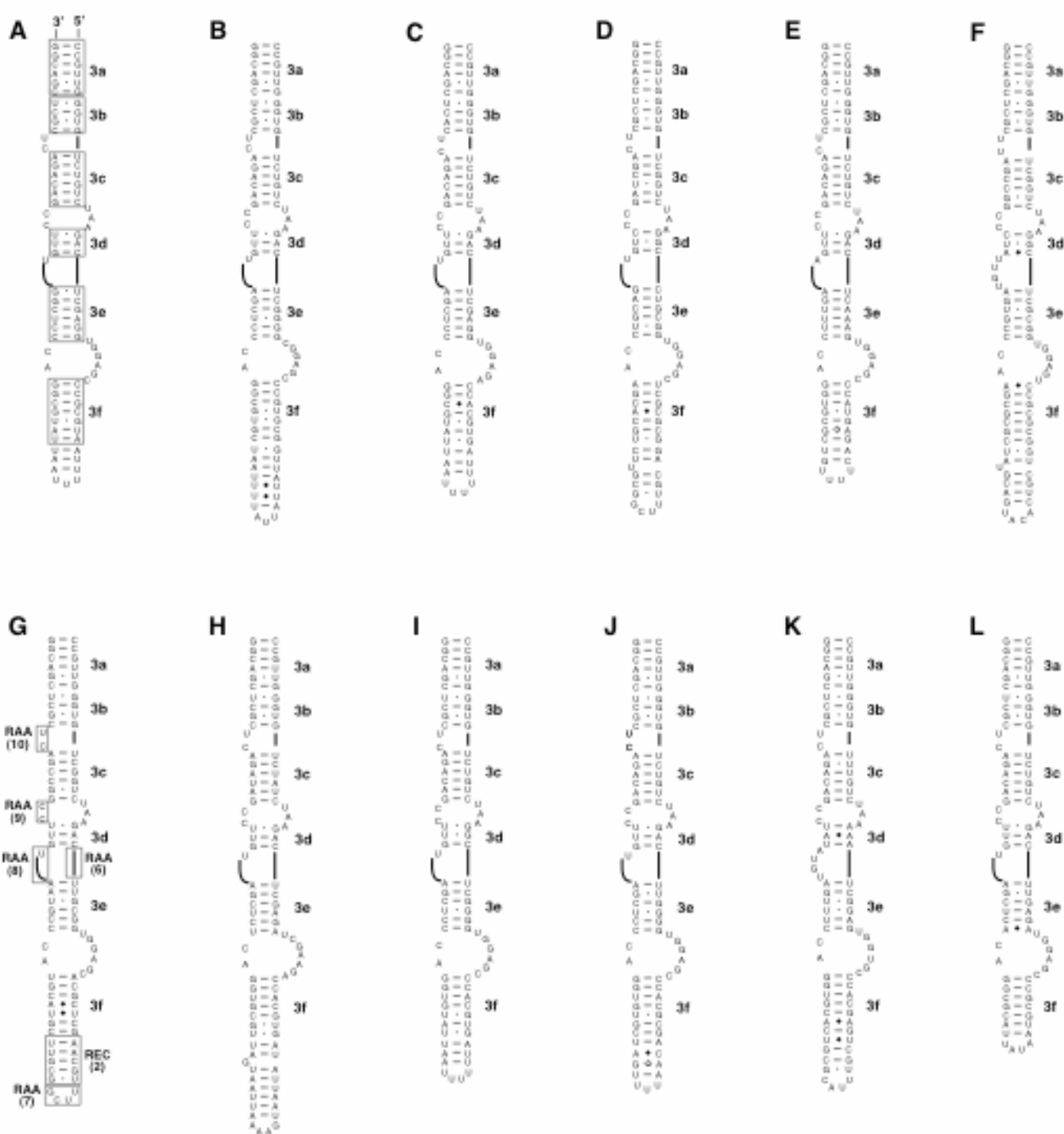


Figure 5. A gallery of diverse secondary structure diagrams from the "helix 3-1" compound helix in the D2 region (synonymous with the 545 gallery of Schnare *et al.* (1996)) is shown for the following chrysomelid taxa: A. *Acalymma vittata*, B. *Agelastica coerulea*, C. *Cerochroa brachialis*, D. *Coptocycla adamantina*, E. *Epitrix fasciata*, F. *Lamprosoma* sp., G. *Metaxyonycha panamensis*, H. *Neolochmaea dilatipennis*, I. *Pyrrhalta aenescens* J. Thailand specimen 11, K. *Walterianella bucki*, L. *Eucerotoma* sp. 344. Notation for the six helical elements is modified from Gillespie *et al.* (2003, 2004). Helices are boxed in A., and ambiguously-aligned regions are boxed in G. The notation for RAAs and RECs is described in Figure 2 and on page 32. The explanations of base-pair symbols and reference for software used to construct structure diagrams are in Figure 3.

The chrysomelid sequence with the largest insertion, *Agelastica coerulea*, has the potential to form an eight base-paired helix in this region (helix **3q** in Fig. 6-A). Other large insertions with different sequences in this region in scarab beetles and apocritan Hymenoptera can form a similar helix (Gillespie, unpubl. data). "Helix 3-2" has five RAAs (Fig. 6F).

Expansion segment D3

The 28S-D3 region, corresponding to the 650 region of the nuclear LSU (Schnare *et al.*, 1996), contains three compound helices in chrysomelids, labeled **D3-1**, **D3-2**, and **D3-3**, following the notation of Kjer *et al.* (2001). In Diptera (Kjer *et al.*, 1994; Schnare *et al.*, 1996; Hwang *et al.*, 1998) and the machilid *Petrobius* sp. (Hwang *et al.*, 1998), the helix **D3-1** is shortened or completely deleted, resulting in only 2 helices (**D3-2** and **D3-3**) in the D3 expansion segment. The basepairs in helix **D3-1** in the chrysomelids are supported by extensive positional covariation for a larger set of sequences that includes the chrysomelids, Trichoptera (Kjer *et al.*, 2001), Odonata (Kjer, pers. comm.) and Hymenoptera (Gillespie, unpubl. data). This suggests that a helix which is present in the other holometabolous insect orders is deleted in Diptera. A gallery of structures representing the three motifs of the D3 in chrysomelids is shown in Figure 7. At least one unpaired nucleotide is flanked by the two helices, **D3-2a** and **D3-2b**. Three RAAs occur in the D3 in chrysomelids (Fig. 7F).

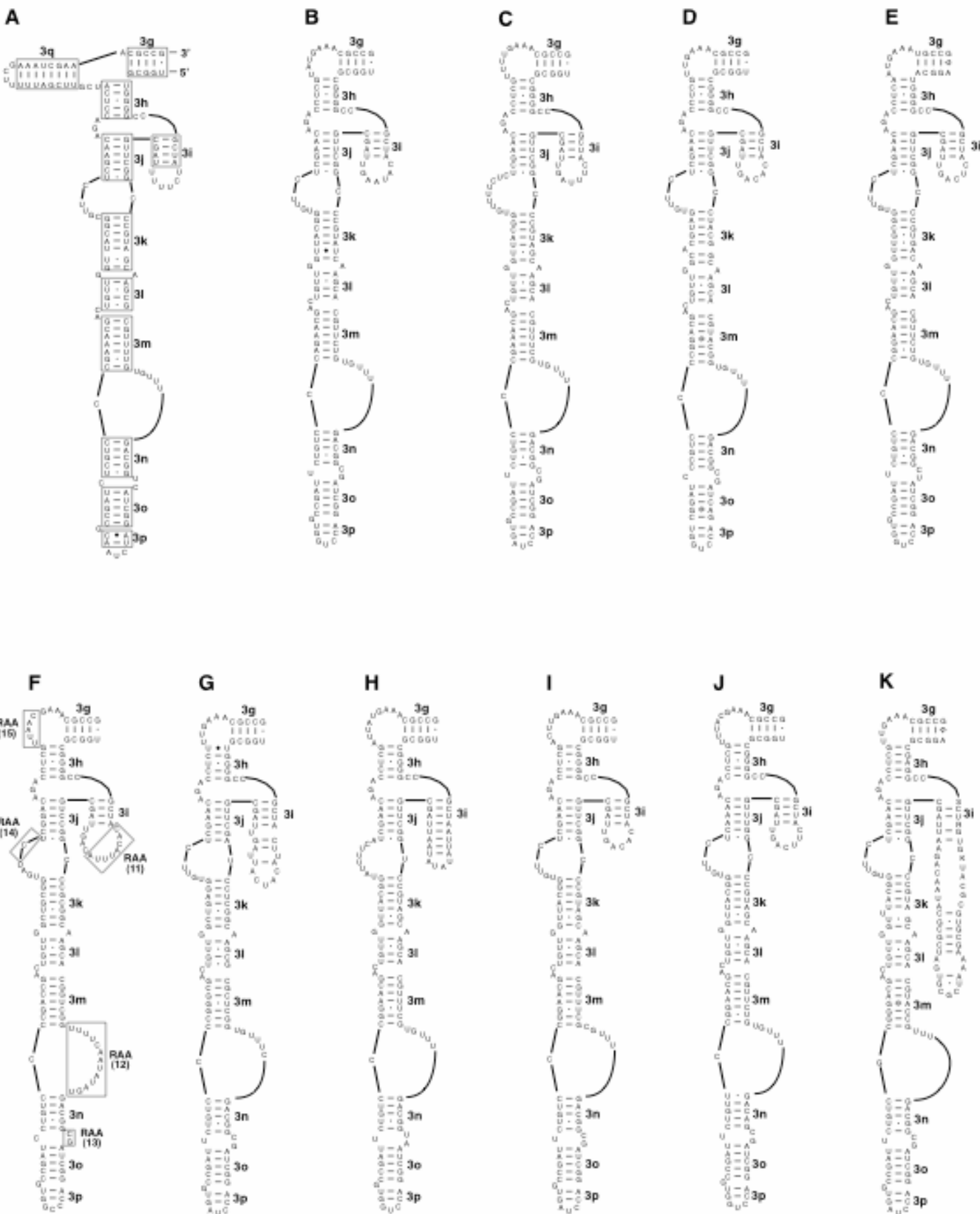


Figure 6. A gallery of diverse secondary structure diagrams from the "helix 3-2" compound helix in the D2 region (synonymous with the 545 gallery of Schnare *et al.* (1996)) is shown for the following chrysomelid taxa: A. *Agelastica coerulea*, B. *Acalymma vittata*, C. *Cerochroa brachialis*, D. *Coptocycla adamantina*, E. *Epitrix fasciata*, F. *Lamprosoma* sp., G. *Metaxyonycha panamensis*, H. *Neolochmaea dilatipennis*, I. *Pyrrhalta aenescens*, J. Thailand specimen 11, K. *Walterianella bucki*. Notation for the 10 helical elements is modified from Gillespie *et al.* (2003, 2004), with the potential base pairing region within RAA (15) in *A. coerulea* named helix 3q. Helices are boxed in A., and ambiguously-aligned regions are boxed in F. The notation for RAAs is described in Figure 2 and on page 32. The explanations of base-pair symbols and reference for software used to construct structure diagrams are in Figure 3.

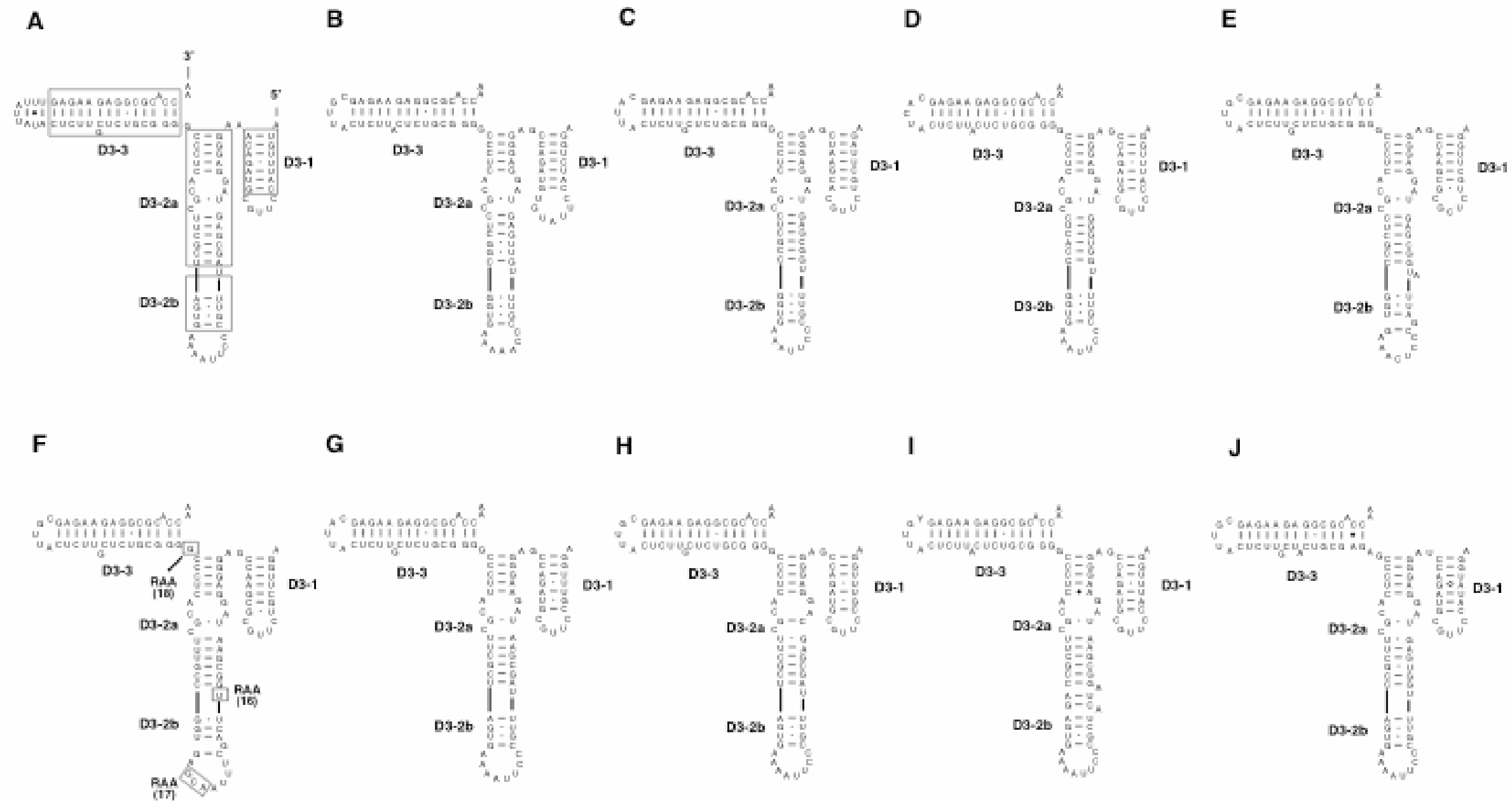


Figure 7. A gallery of diverse secondary structure diagrams for the D3 region (synonymous with the 650 gallery of Schnare *et al.* (1996)) is shown for the following chrysomelid taxa: A. *Cerochroa brachialis*, B. *Scelidopsis* sp., C. *Coptocycla adamantina*, D. *Epitrix fasciata*, E. *Lamprosoma* sp., F. *Metaxyonycha panamensis*, G. *Neolochmaea dilatipennis*, H. *Pyrrhalta aenescens*, I. Thailand specimen 11, J. *Mimastra gracilicornis*. Notation for the 3 compound helices follows the convention of Kjer *et al.* (2001) with the exception of helix **D3-2** being separated into **D3-2a** and **D3-2b**. Helices are boxed in A., and ambiguously-aligned regions are boxed in F. The notation for RAAs is in Figure 2 and on page 32. The explanations of base-pair symbols and reference for software used to construct structure diagrams are in Figure 3.

Core elements

The D2 and D3 expansion regions are flanked by segments of the core rRNA structure. In contrast with the D2 and D3 regions, the core region usually has less insertions and deletions and more sequence conservation. The sequence between D2 and D3, including the 5' and 3' halves of helices H589, H604, H628, H700, and H563, and the 5' half of helices H579, H671 and H687 were determined with the D2 and D3 sequences.

Helical conservation

Characteristic patterns of nucleotide substitutions and positional covariation in the expansion segments D2 and D3 reveal 30 conserved helices in the secondary structure model in the chrysomelids (Table 1). A total of 55.7% of the basepairs within the helical regions of the D2 and D3 chrysomelid expansion segments (not including the core regions sequenced) exhibit some degree of covariation (61.16% in D2, 37.84% in D3; calculated from Table 1). Within the chrysomelid dataset, the more variable positions within helices usually have more positional covariation at a larger percentage of the proposed basepairs, while the positions that are more conserved have a minimal amount of covariation among the two positions that are basepaired. While many of the basepairs in the helices in the D2 and D3 secondary structure model have extensive amounts of positional covariation, some of the sequences underlying the helices, including **2**, **2a**, **3**, **3a**, **3h**, **3l**, **3o**, **3p** and **D3-3**, are conserved within the chrysomelids, and thus have minimal or no comparative support. However, sequence variation between the chrysomelids and other insect taxa D2 and D3 sequences contains positional covariations

that substantiate the proposed basepairs in the structure model (<http://www.rna.icmb.utexas.edu/>). The frequency of the four nucleotides in the unpaired regions of the chrysomelid D2 and D3 sequences is approximately 25% per base, while the paired regions have a bias for guanine (40%) and pyrimidines (46%) (Table 2). This unequal nucleotide frequency can be attributed to the ability of guanine to basepair with both cytosine and uracil (reviewed in Gutell *et al.*, 1994). An analysis of the ratio of transitions to transversions (ts/tv) in paired and unpaired regions reveals a bias for more transitions in paired regions (Table 2). This is consistent with a mutational mechanism under selection for compensatory base changes repairing deleterious substitutions (Wheeler & Honeycutt, 1988; Rousset *et al.*, 1991; Kraus *et al.*, 1992; Gatesy *et al.*, 1994; Vawter & Brown, 1993; Nedbal *et al.*, 1994; Douzery & Catzeflis, 1995; Springer *et al.*, 1995; Springer & Douzery, 1996). While it is expected that transversions should occur in greater frequency than transitions in regions without an expected ts/tv bias (Jukes & Cantor, 1969), such as RNA helices, we interpret a transition bias in non-pairing regions as a consequence of not including the majority of transversions that likely occur in the hypervariable regions wherein nucleotide homology could not be confidently assigned. In summary, our covariation analyses strongly support our predicted model (Fig. 3) for the expansion segments D2 and D3 from these sampled chrysomelid taxa.

Table 1. Composition and degree of compensation for the base pairs of the D2 and D3 expansion segments and related core regions of the 28S rRNA in rootworms and related chrysomelid beetles. For base composition percentages, bold values represent any base pair present at 2% or greater in the alignment. Underscored values show which base pair types strictly covary for that base pair, with the summed underscored numbers providing a percentage of covariation (note: this approach does not account for intermediate GU pairs).

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical								Non-canonical									
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
D2 2	1	168	<u>10.1</u>	0	0	<u>78.0</u>	11.9	0	0	0	0	0	0	0	0	0	0	0	0	Y
	2	167	97.6	0	0	0	1.2	0	0	1.2	0	0	0	0	0	0	0	0	0	Y
	3	173	99.4	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	178	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	5	178	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	6	178	0	0	0	98.9	0.6	0	0	0	0	0	0	0	0	0	0	0	0.6	N
2a	1	196	0	99.0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0	0	0	N
	2	194	95.4	0	0	0	4.1	0	0	0	0	0	0.5	0	0	0	0	0	0	N
	3	196	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	197	99.0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	N
	5	195	0	0	97.9	0	0	0	0	0	0	0	0	0	2.1	0	0	0	0	N
	6	196	0	0	95.4	0	0	0	0	0	0	4.6	0	0	0	0	0	0	0	N
	7	194	0	0	0	0	0	99.5	0	0	0	0	0	0	0	0	0	0	0.5	N
2b	1	192	<u>97.9</u>	0	0	<u>1.0</u>	0	0	0	0.5	0	0	0.5	0	0	0	0	0	0	Y
	2	199	<u>2.0</u>	<u>1.0</u>	<u>0.5</u>	<u>57.8</u>	36.7	0	0	0.5	0	0	0	0	0	0	0	1.5	0	Y
	3	199	0	<u>66.8</u>	<u>8.0</u>	0	0	21.1	0	0	1.0	0.5	0.5	0	0	0	0	2.0	0	Y
2c	1	199	13.6	0	0	4.0	<u>79.4</u>	0.5	0	0	0	0	0	0.5	0	0	<u>0.5</u>	1.5	0	Y
	2	199	0	<u>3.0</u>	<u>88.9</u>	0.5	<u>1.0</u>	5.0	0.5	0	0	0.5	0	0	0	0	0	0.5	0	Y
	3	198	0	<u>87.9</u>	<u>1.5</u>	<u>0.5</u>	0	9.1	0	0	0.5	0	0	0	0	0	0	0	0.5	Y
	4	194	<u>94.8</u>	0	<u>2.1</u>	<u>0.5</u>	1.5	0	0	0.5	0	0	0	0	0	0.5	0	0	0	Y
	5	196	<u>10.7</u>	0	0	<u>82.1</u>	5.6	0	0	0	0	0	0.1	0	0.5	0	0	0	0	Y
2d	1	199	<u>1.5</u>	0	<u>65.8</u>	0.5	0	0.5	5.0	0	0	0	0.5	<u>0.5</u>	1.0	0	0	24.6	0	Y
	2	197	0	4.1	0.5	1.0	0	<u>77.7</u>	0	0	1.0	0	0	<u>3.0</u>	0	6.1	1.0	5.6	0	Y
	3	195	<u>72.8</u>	0	<u>0.5</u>	0	3.6	0	0	17.9	<u>0.5</u>	0	0	0	1.5	1.5	1.0	0	0.5	Y

Table 1 Continued.

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical						Non-canonical											
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
2e	1	198	<u>9.6</u>	0	0	<u>63.1</u>	<u>26.3</u>	<u>0.5</u>	0	0	0	0	0	0	0	0	0	0.5	0	Y
	2	199	<u>0.5</u>	0	0	<u>76.4</u>	<u>22.1</u>	0	0	0	0	0	0	0	0	0.5	0	0.5	0	Y
	3	197	0	<u>58.9</u>	<u>19.8</u>	<u>0.5</u>	0	<u>20.8</u>	0	0	0	0	0	0	0	0	0	0	0	Y
	4	198	<u>43.9</u>	0	0.5	<u>3.5</u>	<u>50.0</u>	0	0	0	0	0	0	0	0.5	0	<u>0.5</u>	1.0	0	Y
	5	198	<u>3.0</u>	<u>1.5</u>	<u>81.8</u>	<u>5.1</u>	<u>2.5</u>	0.5	0	0	0	0	0	0	0	0	0	<u>5.6</u>	0	Y
2f	1	199	0	<u>99.5</u>	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	N
	2	196	<u>55.6</u>	0	0	<u>1.0</u>	<u>42.9</u>	0	0	0	0	0	0.5	0	0	0	0.5	0	0	Y
	3	198	<u>58.1</u>	0	0	<u>21.7</u>	<u>19.2</u>	0	0.5	0	0	0	0	0	0	0	0	0	0.5	Y
	4	200	<u>0.5</u>	0	<u>2.5</u>	<u>89.0</u>	<u>4.5</u>	0	0.5	0.5	0	0	0	0	0	0	0	1.0	1.5	Y
3	1	198	0	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	200	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	201	0	0	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	200	0	<u>98.5</u>	<u>1.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	5	201	<u>99.5</u>	0	0	<u>0.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	6	197	0	<u>85.8</u>	<u>13.7</u>	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	Y
3a	1	203	0	<u>99.5</u>	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	N
	2	203	0	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	203	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	202	0	0	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	5	203	0	0.5	0	0	0	<u>99.5</u>	0	0	0	0	0	0	0	0	0	0	0	N
	6	203	<u>100</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
3b	1	203	0.5	0	0	0.5	<u>99.0</u>	0	0	0	0	0	0	0	0	0	0	0	0	Y
	2	203	<u>99.5</u>	0	0	<u>0.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	3	203	0	<u>3.9</u>	<u>9.9</u>	0	0	<u>83.7</u>	0	0	0	0	0	0	0	0	0	<u>2.5</u>	0	Y
	4	203	<u>96.6</u>	0	0	0	<u>2.5</u>	0	0	1.0	0	0	0	0	0	0	0	0	0	Y
3c	1	203	0	0	<u>99.0</u>	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0	N
	2	203	0	<u>94.6</u>	<u>1.0</u>	0	0	<u>3.4</u>	0	0	0	1.0	0	0	0	0	0	0	0	Y
	3	203	<u>10.3</u>	0	<u>89.7</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	4	203	<u>93.6</u>	0	0	<u>1.0</u>	<u>5.4</u>	0	0	0	0	0	0	0	0	0	0	0	0	Y
	5	203	0	0	<u>90.6</u>	0	0	9.4	0	0	0	0	0	0	0	0	0	0	0	N

Table 1 Continued.

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical						Non-canonical											
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
3d	6	201	0	98.0	0	0	0	2.0	0	0	0	0	0	0	0	0	0	0	0	N
	1	203	31.0	0	0	1.5	66.5	0	0	0	0	0	0	0	0	1.0	0	0	0	Y
	2	203	0	0	0	64.5	34.0	0	1.5	0	0	0	0	0	0	0	0	0	0	N
	3	203	0	79.8	11.8	0.5	0.5	1.5	0	0	1.0	3.0	0	1.0	0	0	0	1.0	0	Y
3e	1	203	0	3.9	73.9	0	0	16.3	0	0	0	0	0	0	0.5	0	0	5.4	0	Y
	2	203	0.5	75.9	3.9	0	0	17.7	0	0	1.5	0	0	0	0	0.5	0	0	0	Y
	3	203	56.7	0	0.5	3.0	38.9	0.5	0	0.5	0	0	0	0	0	0	0	0	0	Y
	4	203	1.5	7.4	0	72.4	16.3	1.0	0	0	0	0	0	0	0	0	0	1.5	0	Y
	5	203	86.2	0.5	0	10.8	2.5	0	0	0	0	0	0	0	0	0	0	0	0	Y
	6	203	89.2	0	1.0	0.5	0	0	8.9	0	0	0	0	0	0.5	0	0	0	0	Y
3f	1	201	0	85.6	2.0	4.5	0	0	0	0	0	8.0	0	0	0	0	0	0	0	Y
	2	202	0	99.5	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	N
	3	203	39.9	0	0	46.3	11.8	0	0	1.5	0.5	0	0	0	0	0	0	0	0	Y
	4	203	0	81.8	1.0	0	0	8.9	0	0	0	7.4	0	0.5	0.5	0	0	0	0	Y
	5	203	46.8	0.5	0	3.0	46.8	0	0	0	0	0	0	0	0	0	0	3.0	0	Y
	6	202	0	29.2	51.5	0	0	14.9	1.5	0	2.0	0	0	0	0	0	0	1.0	0	Y
	7	201	30.3	0	0	39.8	28.4	0	0	0.5	0	0	0	0	0.5	0	0	0.5	0	Y
3g	1	202	0	1.5	2.5	0	0	89.6	1.0	0	5.4	0	0	0	0	0	0	0	0	Y
	2	203	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	201	99.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	N
	4	202	0	97.5	2.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	Y
	5	203	98.0	0	0	1.0	0.5	0	0	0	0	0	0	0	0	0	0	0	0.5	Y
3h	1	202	0	86.6	7.4	1.0	0	0	0	0	0	0.5	0.5	0	0	0	0	3.5	0.5	Y
	2	203	96.6	0	0	1.5	0.5	0	0	1.5	0	0	0	0	0	0	0	0	0	Y
	3	203	1.5	0	0	29.1	69.5	0	0	0	0	0	0	0	0	0	0	0	0	Y
	4	202	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	5	203	99.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	N
3i	1	202	99.5	0	0.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	2	201	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N

Table 1 Continued.

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical						Non-canonical											
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
3j	3	203	0	0	<u>99.5</u>	<u>0.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	4	202	0	0	0	<u>99.5</u>	0.5	0	0	0	0	0	0	0	0	0	0	0	0	N
	1	202	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	203	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	203	0	<u>1.5</u>	<u>98.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	4	203	0	<u>97.5</u>	<u>2.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
3k	5	203	99.5	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	N
	6	203	0	0	0	4.9	95.1	0	0	0	0	0	0	0	0	0	0	0	0	N
	1	203	0	<u>92.6</u>	<u>3.4</u>	0	0	0.5	0	0	1.0	1.0	0	0	0	0	0	0	1.5	Y
	2	203	0	<u>98.5</u>	<u>1.0</u>	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	Y
	3	202	<u>95.0</u>	0	<u>3.0</u>	<u>1.5</u>	0.5	0	0	0	0	0	0	0	0	0	0	0	0	Y
	4	203	0	<u>9.4</u>	<u>67.0</u>	0	0	23.2	0	0	0	0	0.5	0	0	0	0	0	0	Y
3l	5	203	<u>6.9</u>	<u>0.5</u>	0	<u>87.2</u>	4.4	0	0	0.5	0.5	0	0	0	0	0	0	0	0	Y
	6	203	11.3	0	0	1.0	82.3	0	0	0	0	0	0	0	0	0	0	5.4	0	Y
	7	202	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	1	202	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	203	0	0	0	0.5	99.5	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	203	0	98.5	0	0	0	0.5	0	0	1.0	0	0	0	0	0	0	0	0	N
3m	4	203	0	0	0	74.9	25.1	0	0	0	0	0	0	0	0	0	0	0	0	N
	1	203	0	<u>97.5</u>	<u>2.0</u>	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0	Y
	2	203	<u>92.1</u>	0	0	<u>0.5</u>	7.4	0	0	0	0	0	0	0	0	0	0	0	0	Y
	3	203	<u>0.5</u>	<u>3.0</u>	<u>90.6</u>	0	0	5.4	0	0	0	0	0	0	0	0	0	0.5	0	Y
	4	203	0	1.5	<u>90.1</u>	0	0	5.9	0	0	<u>2.5</u>	0	0	0	0	0	0	0	0	Y
	5	202	0	<u>75.7</u>	<u>7.4</u>	0	0	15.8	0	0	0	0	1.0	0	0	0	0	0	0	Y
3n	6	203	<u>10.3</u>	<u>24.1</u>	<u>35.5</u>	0	0	30.0	0	0	0	0	0	0	0	0	0	0	0	Y
	7	203	<u>99.0</u>	0	0	0	0	<u>0.5</u>	0	0	0	0	0	0	0.5	0	0	0	0	Y
	1	203	<u>93.1</u>	0	0	<u>0.5</u>	5.9	0	0	0	0	0	0.5	0	0	0	0	0	0	Y
	2	203	<u>0.5</u>	0	0	<u>99.5</u>	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	3	203	0	<u>89.7</u>	<u>1.5</u>	0	0	8.9	0	0	0	0	0	0	0	0	0	0	0	Y
	4	203	4.4	0	0	10.3	85.2	0	0	0	0	0	0	0	0	0	0	0	0	Y

Table 1 Continued.

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical								Non-canonical									
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
3o	5	203	99.0	0	0	0	1.0	0	0	0	0	0	0	0	0	0	0	0	0	N
	1	203	0	0	0	99.0	0.5	0	0	0	0	0	0	0.5	0	0	0	0	0	N
	2	203	0	0	93.6	0	0	6.4	0	0	0	0	0	0	0	0	0	0	0	N
	3	203	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	202	97.5	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	1.5	Y
	5	202	94.1	0	0	0	5.9	0	0	0	0	0	0	0	0	0	0	0	0	N
3p	1	201	0	0	0	97.0	0	0	0	2.5	0	0	0	0.5	0	0	0	0	0	N
	2	202	0	97.5	2.0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	Y
Core H88	1	161	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	161	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	161	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	161	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
27	1	138	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	141	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	N
	3	141	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	141	99.3	0	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	N
	5	141	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	6	142	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	N
	7	142	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	8	142	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	9	142	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	10	142	1.4	0	0	0	98.6	0	0	0	0	0	0	0	0	0	0	0	0	N
	11	144	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	12	144	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	13	144	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
28	1	152	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	152	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	N
	3	152	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	152	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	N

Table 1 Continued.

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical						Non-canonical											
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
	5	152	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	6	153	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	7	153	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	N
	8	153	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	9	153	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
29	1	152	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	152	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	N
D3																				
D3-1	1	151	99.3	0	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	2	151	94.7	0	0	2.0	3.3	0	0	0	0	0	0	0	0	0	0	0	0	Y
	3	151	0	0	99.3	0	0	0.7	0	0	0	0	0	0	0	0	0	0	0	N
	4	151	0	3.3	9.9	0	0	85.4	0	0	1.3	0	0	0	0	0	0	0	0	Y
	5	152	0	9.2	77.0	11.2	0	2.6	0	0	0	0	0	0	0	0	0	0	0	Y
	6	152	9.2	0	0	86.2	4.6	0	0	0	0	0	0	0	0	0	0	0	0	Y
	7	152	0	94.7	1.3	0	0	3.9	0	0	0	0	0	0	0	0	0	0	0	Y
D3-2a	1	148	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	149	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	149	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	149	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	5	149	85.2	0	0	1.3	0	0	0	13.4	0	0	0	0	0	0	0	0	0	Y
	6	149	0	2.0	0	0	0	98.0	0	0	0	0	0	0	0	0	0	0	0	N
	7	148	65.5	0	0	0	0	0	0	34.5	0	0	0	0	0	0	0	0	0	N
	8	149	1.3	0	0	92.6	6.0	0	0	0	0	0	0	0	0	0	0	0	0	Y
	9	148	97.3	0	0	0	2.7	0	0	0	0	0	0	0	0	0	0	0	0	N
	10	150	0	92.7	3.3	0	0	4.0	0	0	0	0	0	0	0	0	0	0	0	Y
	11	149	97.3	0	0	1.3	0.7	0.7	0	0	0	0	0	0	0	0	0	0	0	Y
	12	150	75.3	0	0	22.0	2.7	0	0	0	0	0	0	0	0	0	0	0	0	Y
D3-2b	1	149	0	0.7	40.9	0	0	55.7	0	0	0	0	0	0	0	0	0.7	2.0	0	N
	2	150	0	14.0	16.0	0	0	67.3	0	0	0	0	0	0	0	0	0	2.7	0	Y
	3	150	2.0	0	0	4.7	90.7	0	0	0	0	0	0	2.0	0	0	0	0	7	Y
	4	150	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N

Table 1 Continued.

Helix ^a	Base pair ^b	No. of sequences compared ^c	Base pair composition, % ^d																Gap ^e (-)	Covarying base pair ^f Y/N
			Canonical						Non-canonical											
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU		
D3-3	1	144	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	144	54.9	0	0	25.7	18.8	0	0	0.7	0	0	0	0	0	0	0	0	0	Y
	3	144	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	144	0	75.0	0	0	0	25.0	0	0	0	0	0	0	0	0	0	0	0	N
	5	144	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	6	144	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	N
	7	144	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	8	144	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	9	144	0	97.2	0	0	0	0	0	0	0	2.8	0	0	0	0	0	0	0	N
	10	146	0	0	99.3	0	0	0	0.7	0	0	0	0	0	0	0	0	0	0	N
	11	146	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	12	146	0	99.3	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	0	N
	13	146	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	14	145	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
Core 34	1	128	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	2	128	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	3	129	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	4	128	0	98.4	0	0	0	1.6	0	0	0	0	0	0	0	0	0	0	0	N
	5	128	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	N
	6	129	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	7	128	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	8	129	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	9	126	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	10	129	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N
	11	128	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	N

^a Helix numbering refers the nucleotide positions shown in Figure 2.

^b Base pairs are numbered from 5'-end of 5'-strand of each helix.

^c Numbers vary at each position due to missing data (?), deletions (-) and possible presence of IUPAC-IUB ambiguity codes.

^d The first nucleotide is that in the 5'-strand.

^e Gaps represent single insertion or deletion events, not indels.

^f A covarying position is defined as having substitutions on both sides of the helix across the alignment.

Table 2. Mean percent nucleotides and mean transition/transversion ratios in pairing (stems) and non-pairing (loops) regions of the D2 and D3 expansion segments of the 28S LSU gene of chrysomelids ^{a b c}.

	Nucleotide composition (%)				Substitutions (Ts/Tv) ^b
	A	C	G	U	
stems	0.15	0.24	0.39	0.22	3.66
loops	0.25	0.25	0.26	0.24	2.30

^a Calculated in MacClade 4.0 (Maddison & Maddison, 2000).

^b Missing data and gaps not included in calculations.

^c Nucleotides within RAAs, RSCs and RECs were not included in calculations.

Regions of ambiguous alignment (RAA)

Positional nucleotide homology could not be confidently assigned to 21 regions of our multiple sequence alignment (Table 3). Eighteen of these unalignable regions are defined as RAA, wherein single insertion and deletion events cannot be assessed as homologous characters across all of the sequences in the alignment, and consistent positional covariation (base-pairing) is not found. Without secondary structure basepairing to guide the establishment of columnar homology in regions with many insertions and deletions (Kjer, 1995; Hickson *et al.*, 1996; Kjer, 1997), we did not establish homology statements within RAAs. These nucleotides in the alignment were contained within brackets and were justified to the left (5'-strand) or right (3'-strand). Within the RAA regions, gaps do not represent insertion and deletion events as they do in the unambiguously aligned data. Instead they represent size variation within each RAA.

Regions of slipped-strand compensation (RSC)

The sequence alignment in one region in the D2 expansion segment cannot be aligned with high confidence due to the inconsistent basepairing in its helix (Table 3). This helix is flanked on both sides by conserved basepairs wherein positional homology assessment is unambiguous. Patterns of covariation were used to confirm inconsistent basepairing across the alignment within this RSC, as suggested by Gillespie (2004). As with RAAs, nucleotides in RSCs were bracketed and aligned to approximate homologous basepairs (when basepairs are proposed) or left or right justified, with gaps inserted to adjust for length heterogeneity as in the RAA regions (see above).

Table 3. A list of the 18 regions of alignment ambiguity (**RAA**), one region of slipped-strand compensation (**RSC**) and two regions of expansion and contraction (**REC**) created in the multiple sequence alignment of the expansion segments D2 and D3 of the 28S LSU rRNA from 229 sampled chrysomelids.

Ambiguous region	length ^a (nts)	non-homologous position ^b	General comments
RAA (1)	0-3	24-25	forms a bulge between strands 2b and 2c
RAA (2)	0-2	40-41	forms a bulge between strands 2e and RSC (1)
RSC (1)	7-8	40-41	assignment of homology unclear due to <i>Acalymma</i> spp. <i>sensu stricto</i> (Gouldi group) forming a different structure, as well as other taxa having unique pairing potentials
RSC (1')	5-6	49-50	deletion in RSC (1') causes a slip in the base-pairing in 9 sampled species on <i>Acalymma</i> s.s. that results in a different structure
REC (1)	5-15	44-45	REC (2) and its complement REC (2') form a hairpin-stem loop that is an extension of helix 2f ; from 5 to 14 base-pairings occur across alignment with lateral and internal bulges present that make the region up to 15 positions in length
REC (1')	5-18	44-45	REC (2') and its complement REC (2) form a hairpin-stem loop that is an extension of helix 2f ; from 5 to 14 base-pairings occur

Table 3 Continued.

Ambiguous region	length ^a (nts)	non-homologous position ^b	General comments
			across alignment with internal bulges present that make the region up to 18 positions in length
RAA (3)	3-5	44-45	non-pairing terminal bulge formed by hairpin-stem loop REC (1) ; motif YYR highly common when 4 nts present
RAA (4)	2-6	54-55	forms a lateral bulge between strands 2e' and 2d'
RAA (5)	0-4	57-58	forms a lateral bulge between strands 2d' and 2c'
RAA (6)	0-3	126-127	along with RAA (8) , forms a internal bulge between helices 3d and 3e
REC (2)	0-8	149-150	REC (2) and its complement REC (2') form a hairpin-stem loop that is an extension of helix 3f ; from 0 to 6 base-pairings occur across the alignment with lateral and internal bulges present that make the region up to 8 positions in length; some taxa have no extension of helix 3f
REC (2')	0-8	149-150	REC (2') and its complement REC (2) form a hairpin-stem loop that is an extension of helix 3f ; from 0 to 6 base-pairings occur across the alignment with lateral and internal bulges present that make the region up to 8 positions in length; some taxa have no extension of helix 3f
RAA (7)	3-5	149-150	non-pairing terminal bulge formed by hairpin-stem loop REC (2) or helix 3f ;
RAA (8)	0-4	170-171	along with RAA (6) , forms an internal bulge between helices 3e and 3d
RAA (9)	2-3	174-175	along with positions 121-123, forms an an internal bulge between helices 3c and 3d
RAA (10)	2-3	180-181	forms a lateral bulge between strands 3c' and 3b'
RAA (11)	0-13	242-243	part of the highly variable terminal loop formed by hairpin-stem 3i ;
RAA (12)	2-13	276-277	forms a highly variable lateral bulge between strands 3m and 3n
RAA (13)	2-4	281-282	along with position 305, forms an internal bulge between helices 3n and 3o
RAA (14)	1-7	336-337	(+ 4 nts 3' to 3k') along with position 254, forms an internal bulge between helices 3j and 3k
RAA (15)	0-20	352-353	highly variable unpaired strand joining the 3' strand with the highly conserved 3 helix; forms helix 3q in <i>Agelastica coerulea</i>

Table 3 Continued.

Ambiguous region	length ^a (nts)	non-homologous position ^b	General comments
RAA (16)	1-4	522-523	forms a lateral bulge separating D3-2a and D3-2b
RAA (17)	0-3	528-529	part of the variable terminal loop formed by helix D3-2b
RAA (18)	1-2	557-558	junction between D3-2a and D3-3 ; AG motif in <i>Mimastra gracilicornis</i> causes ambiguous alignment of Gs and As; most likely 1 nt long

^a Refers to the range of nucleotides within each ambiguous region.

^b Nucleotide positions flanking ambiguous regions are given in Figure 2.

Underlined positions represent structures that are not consistent across the alignment (Fig. 2).

Regions of expansion and contraction (REC)

The sequence alignment in two other helical regions in the D2 expansion segment also cannot be aligned with high confidence due to the inconsistent basepairing in their helices (Table 3). Both of these regions have variation in the length of the terminal helix in compound helices "helix 2" and "helix 3-1", thus the precise placement of nucleotides and indels in the alignment is uncertain. While consistent homology statements could not be made in these two ambiguous regions across all sequences in the alignment, secondary structure basepairing was used to differentiate between the helical component and the terminal bulge that comprise the entire hairpin-stem loop structure (see Gillespie, 2004). After bracketing, nucleotides in RECs were treated the same as RSCs (see above).

Taxonomic implications

Structural characters that are unique and characteristic for the tribes, subtribes, sections, and genera of the Luperini were identified (Table 4). These signatures in the D2 and D3 regions are consistent with previous taxonomic delineations within the Galerucinae *s.s.* (Leng, 1920; Laboisrière, 1921; Weise, 1923; Wilcox, 1965; Seeno & Wilcox, 1982). The majority of taxon-specific structural characters in these molecules are located in the hairpin-stem loops of helices **2f** and **3f**. A more detailed depiction of these taxon-specific structural characters superimposed over our multiple sequence alignment is posted at <http://hisl.tamu.edu>. Individual secondary structure diagrams are also available (see below) that illustrate taxon-specific structural characters defined by our alignment. Calculated nucleotide frequencies for each higher-level taxon indicate that there are no significant differences between any of the sampled taxa regarding the distribution of the four bases throughout this region of the 28S (data not shown).

Utility for phylogeny reconstruction

The alignment of rDNA sequences becomes progressively more difficult as the sequence and length variation increases. The accuracy of the phylogenetic reconstruction is dependent in part on the accuracy of the alignment of the rDNA sequences. The expansion segments of the eukaryotic LSU rRNA are unique because they accumulate an extreme amount of nucleotide insertions (Veldman *et al.*, 1981; Michot *et al.*, 1984), and yet presumably have little impact on the function of the ribosome in translation (Musters *et al.*, 1989; Sweeney & Yao, 1989; Musters *et al.*, 1991), with the exception of

Table 4. Secondary structure characters of the D2, D3 expansion segments from the higher-level chrysomelid taxa sampled in this analysis. General comments describe the conservation of these characters, and whether or not they are found in unrelated taxa.

Taxon	Region^a	Character^b	General Comments
<i>Dircema</i> spp.	RAA (2)	GU	Internal bulge absent except for CC in <i>Lamprosoma</i> and single insertions in 3 flea beetles
<i>Acalymma</i> spp. s.s.	RSC (1)	C-UCUU	Deletion causes slippage in the hydrogen-bonding in this region that differs from the rest of the taxa in the alignment
-----	RSC (1')	variable	Helix 2f expands and contracts across the alignment with positional homology uncertain; base composition in this helix, as well as sequence length, defines many genera and subtribes of the Luperini
<i>Dircema</i> spp.	RAA (3)	UUU	Triloop formed by extended 2f helix; UCG in <i>Aplosomyx quadripustulatus</i> and <i>Mimastra gracilicornis</i> ; usually a tetraloop with a conserved UUYG motif
Galerucinae s.s.	RAA (5)	R	Single base-pair internal bulge is variable outside of the strict subfamily; U in <i>Medythia suturalis</i>
-----	REC (2)	variable	Helix 3f expands and contracts across the alignment with positional homology uncertain; base composition in this helix, as well as sequence length, defines many genera and subtribes of the Luperini
-----	RAA (3)	UUU	Triloop formed by extended 3f helix; base composition in this loop, as well as sequence length, defines many genera and subtribes of the Luperini, as well as generic groups in other chrysomelid subfamilies; loop is consistently larger in non-galerucine taxa
Oedionychina	pos. 213-239	large insert	These 3 flea beetles have an insertion within the terminal loop formed by helix 3i
-----	RAA (11)	variable	Terminal loop formed by helix 3i is informative at the generic level; however, certain motifs, such as CUU, are homoplastic
<i>Agelastica coerulea</i>	RAA (15)	8-bp helix	The ambiguous region between strands 3h' and 3g' forms a stable helix (helix 3q); may be a common insertion site as helices form here in other insects

^a Regions within the D2 and D3 can be found in Figure 2.

^b Illustration of structural characters can be found at <http://hisl.tamu.edu/>

expansion segment D8, which is thought to interact with small nucleolar RNA E2 (Rimoldi *et al.*, 1993, Sweeney *et al.*, 1994). Extraordinary differences in sequence length (Gutell, 1992; De Rijk *et al.*, 1994) and secondary structure in expansion segments, even in recently-diverged organisms, are not uncommon (Hillis & Dixon, 1991; Schnare *et al.*, 1996; Gillespie, unpubl. data). Thus, severe deviations from a common structure in eukaryotic expansion segments are expected (Schnare *et al.*, 1996), especially among taxa that have diverged over a large evolutionary time scale.

While seemingly problematic, the above characteristics of the expansion segments of the nuclear LSU rRNA make these markers ideal for phylogeny reconstruction. Conserved regions involved in hydrogen-bonding can be used to delimit regions wherein primary assignment of homology is uncertain and indefensible (Kjer, 1997; Lutzoni *et al.*, 2000; Kjer *et al.*, 2001). The assignment of positional homology in length heterogeneous datasets based on biological criteria has been shown to improve phylogeny estimation (Dixon & Hillis, 1993; Kjer, 1995; Titus & Frost, 1996; Morrison & Ellis, 1997; Uchida *et al.*, 1998; Mugridge *et al.*, 1999; Cunningham *et al.*, 2000; Gonzalez & Labarere, 2000; Hwang & Kim, 2000; Lydeard *et al.*, 2000; Morin, 2000; Xia, 2000; Xia *et al.*, 2003). Recoding RAAs and RECs as complex multistate characters with (Lutzoni *et al.*, 2000; Xia *et al.*, 2003; Gillespie *et al.*, 2003, 2004) or without (Kjer *et al.*, 2001; Gillespie *et al.*, 2003, 2004) the implementation of an unequivocal weighting scheme can retain phylogenetic information in these unalignable regions. Also, the descriptive coding of unalignable positions as morphological characters based on secondary structure can extract information from these regions of

rRNA in phylogenetic analysis (Billoud *et al.*, 2000; Collins *et al.*, 2000; Lydeard *et al.*, 2000; Ouvard *et al.*, 2000; Gillespie, unpubl. data).

Model applicability

Unpublished data from our labs suggest that the structural model presented here for the D2 and D3 expansion segments of the 28S rRNA gene from chrysomelids is applicable for several insect groups, including ichneumonoid, chalcidoid, proctotrupoid and cynipoid Hymenoptera, scaraboid and curculionoid Coleoptera, and lower level studies on adephagous and other polyphagous beetles, including cassidine Chrysomelidae. All of these insect lineages contain the seven compound helices described in our model, with the majority of the length and structure variation occurring in the most distal regions of these compound helices (Gillespie, unpubl. data). Our model is consistent with the predicted structure of the *Drosophila melanogaster* D2 region (Schnare *et al.*, 1996). The only significant difference is a reduced "helix 3-2" in the fruit fly (helix K in Schnare *et al.* (1996)). Interestingly, predicted D2 structures for the plant *Arabidopsis thaliana*, the fungus *Cryptococcus neoformans*, and the protist *Chlorella ellipsoidea* also share the general 4-compound helix model presented here, but contain minor differences in the size of "3-1 helix" and "3-2 helix" and the length of the unpaired regions linking these motifs to the highly conserved helices **3a** and **3** (synonymous with helix H2 of Michot & Bachellerie (1987)). These structural similarities between highly divergent taxa may suggest that similar regions of the D2 have the propensity to expand and contract over time, possibly a consequence of mild structural conservation that limits

mutations to these specific locations. These findings are consistent with those of Wuyts *et al.* (2000) for the variable region 4 (V4) of the SSU rRNA across eukaryotes. Lower level studies of mitochondrial rRNA from Odonata (Misof & Fleck, 2003) and Phthiraptera (Page *et al.*, 2002) also support this phenomenon of helix birth and death across divergent lineages.

Given the relative conservation within these variable regions of the 28S rRNA, the establishment of primary nucleotide homology across insects may be possible for some groups, particularly those within the Holometabola. However, with increased sequence divergence, it is likely that many regions of the D2 and D3 expansion segments will prove unalignable and non-comparable at the nucleotide level. For instance, published structural models for the expansion segment D3 from Diptera suggest severe deviations from the 3 compound helices defined by our model (Hancock *et al.*, 1988; Tautz *et al.*, 1988; Schnare *et al.*, 1996; Hwang *et al.*, 1998). This could possibly be the result of an accelerated rate of nucleotide substitution that presumably occurred in basal lineages of Diptera (Friedrich & Tautz, 1997). This is supported in part by our D3 model, and the D3 model for Amphiesmenoptera (Kjer *et al.*, 2001) and Odonata (Kjer, pers. comm.), which are more consistent with chordate and nematode D3 structures (compiled in Schnare *et al.*, 1996) than those of Diptera (Hancock *et al.*, 1988; Tautz *et al.*, 1988; Schnare *et al.*, 1996; Hwang *et al.*, 1998). This accelerated substitution rate, however, does not explain why the D2 is so structurally different in lower Diptera (Nematocera) than in derived flies (Brachycera), as our D2 model is not congruent with any structural predictions for this region in *Aedes albopictus* (Kjer, *et al.* 1994; Schnare

et al., 1996). Interestingly, our model and these published dipteran models, are quite different than preliminary structures of Strepsipteran D2 (Gillespie, unpubl. data) and D3 (Hwang *et al.*, 1998) expansion segments.

Experimental procedures

Taxa examined

Table 5 lists the chrysomeloid species analyzed in this investigation, with respective GenBank accession numbers for all sequences given. For the 28S-D2 we combined 65 new sequences with 137 from a previous study (Gillespie *et al.*, 2004). The 153 sequences of the 28S-D3 segment were generated in this investigation. All 229 taxa are represented by the 28S-D2 region, with 50 taxa missing the 28S-D3 expansion segment. Voucher specimens for all sampled taxa can be found in the Texas A&M University, Rutgers University, or the University of Delaware insect museums. Information regarding sampled taxa is available at the following website (<http://hisl.tamu.edu>).

Genome isolation, PCR, and sequencing

For the sequences generated in this study, total genomic DNA was isolated using DNeasy™ Tissue Kits (Qiagen). PCR conditions followed those of Cognato & Vogler (2001), with primers designed for amplification of both the D2 and D3 expansion segments found in Gillespie *et al.* (2003, 2004). Double-stranded DNA amplification products were sequenced directly with ABI PRISM™ (Perkin-Elmer) Big Dye Terminator Cycle Sequencing Kits and analyzed on an Applied Biosystems (Perkin-

Table 5. A list of the chrysomeloid taxa analyzed in this investigation.

Taxon^a (Family/Subfamily/Tribe/Subtribe/Section)	Extract Code^b	Accession Number
Orsodacnidae		
<i>Orsodacne atra</i> (Ahrens)	JJG114	AY243660
^K <i>Orsodacne atra</i> (Ahrens)	CND114	AY171422
Chrysomelidae		
Lamprosomatinae		
<i>Lamprosoma</i> sp. Kirby	JJG215	AY243651
Clytrinae		
<i>Cyltrasoma palliatum</i>	JJG286	AY646286
Criocerinae		
<i>Lema</i> sp. Fabricius	JJG308	AY243659
Cassidinae		
<i>Coptocycla adamantina</i> (Germar)	JJG214	AY243649
<i>Microrhopala vittata</i> Baly	JJG218	AY243650
Eumolpinae		
<i>Syneta</i> sp.	CND723	AY646287
^K <i>Syneta adamsi</i> Baly	SJK723	AY171441
<i>Megascelis</i> sp. Latreille	JJG244	AY243652
<i>Metaxyonycha panamensis</i> Jacoby	JJG311	AY646288
<i>Metaxyonycha</i> sp. Chevrolat	JJG132	AY243653
<i>Callisina quadripustulata</i> Baly	JJG321	AY243654
<i>Colaspis</i> sp. Fabricius (or nr.)	JJG357	AY646289
<i>Colaspis</i> sp. Fabricius	JJG141	AY243655
<i>Colasposoma</i> sp. Laporte	JJG318	AY243656
<i>Tymnes tricolor</i> (Fabricius)	JJG258	AY243657
<i>Chalcophana</i> sp. Chevrolat	JJG352	AY243658
Chrysomelinae		
Chrysomelini		
<i>Chrysomela knabi</i> Brown	JJG237	AY243661
<i>Chrysomela aeneicollis</i> (Schaeffer)	JJG277	AY243662
<i>Chrysomela populi</i> Linnaeus	JJG236	AY243663
^K <i>Chrysomela tremulae</i> Fabricius	SJK705	AY171423
^K <i>Chrysolina coerulans</i> (Scriba)	SJK703	AY171429
<i>Gastrophysa cyanea</i> Melsheimer	JJG329	AY243664
^K <i>Paropsis porosa</i> Erichson	SJK704	AY171438
^K <i>Zygogramma piceicollis</i> (Stål)	CND334	AY171440
Timarchini		
<i>Timarcha</i> sp. Latreille	CND706	AY646290
^K <i>Timarcha tenebricosa</i> (Fabricius)	SJK707	AY171439
Galerucinae sensu lato		
Alticini		
^K <i>Altica</i> sp. Geoffroy	CND221	AY171424
^K <i>Allochroma</i> sp. Clark	CND327	AY171428
^K <i>Aphthona nigriscutis</i> Foudras	SJK700	AY171430
^K <i>Chaetocnema</i> sp. (Stephens) (nr. <i>costulata</i>)	SJK720	AY171431
^K <i>Disonycha conjuncta</i> (Germar)	CND061	AY171434
^K <i>Blepharida rhois</i> (Forster)	CND209	AY171435
^K <i>Dibolia borealis</i> Chevrolat	CND419	AY171442
^K <i>Sangariola fortunei</i> (Baly)	SJK721	AY171443
<i>Systema</i> sp. Chevrolat (nr. <i>lustrans</i>)	JJG219	AY243665
^K <i>Systema bifasciata</i> Jacoby	SJK219	AY171432
<i>Scelidopsis</i> sp. Jacoby	JJG225	AY243666
<i>Cacoscelis</i> sp. Chevrolat	JJG195	AY243667

Table 5 Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section)	Extract Code ^b	Accession Number
<i>Epitrix fasciata</i> Blatchley	JJG328	AY243668
<i>Physodactyla rubiginosa</i> (Gerstaecker)	CND253	AY243671
<i>Alagoasa libentina</i> (Germar)	CND303	AY243670
<i>Walterianella bucki</i> Bechyné	CND039	AY243673
<i>Blepharida ornata</i> Baly	CND209	AY243672
<i>Megistops vandepolli</i> Duvivier	CND002	AY243669
<i>Luperaltica</i> sp. Crotch (or nr.)	JJG253	AY243695
^K <i>Orthaltica copalina</i> (Fabricius)	SJK721	AY171437
<i>Aedmon morrisoni</i> Blake	CND207	AY646291
Galerucinae sensu stricto		
Oidini		
<i>Oides decempunctata</i> (Billberg)	JJG334	AY243674
^K <i>Oides decempunctata</i> (Billberg)	SJK718	AY171448
<i>Oides andrewsi</i> Jacoby	JJG409	AY646292
<i>Oides andrewsi</i> Jacoby	JJG439	AY646293
<i>Anoides</i> sp. Weise (or nr.)	JJG380	AY646294
Galerucini		
Galerucini Chapuis "genus undet."	JJG387	AY646295
Galerucites		
<i>Galeruca</i> sp. Geoffroy	CND700	AY646297
^K <i>Galeruca rudis</i> LeConte	CND702	AY171436
Coelomerites		
<i>Caraguata pallida</i> (Jacoby) (or nr.)	JJG139	AY243776
<i>Dircema cyanipenne</i> Bechyné (or nr.)	JJG118	AY243771
<i>Dircema</i> sp. Clark	JJG343	AY243772
<i>Dircema</i> sp. Clark (or nr.)	JJG350	AY646298
<i>Dircema</i> sp. Clark	JJG355	AY646299
<i>Dircema</i> sp. Clark	JJG449	AY646300
<i>Dircemella</i> sp. Weise	JJG202	AY243773
<i>Dircemella</i> sp. Weise	JJG307	AY243774
<i>Trirhabda bacharidis</i> (Weber)	JJG075	AY243769
^K <i>Monocesta</i> sp. Clark	CND710	AY171433
<i>Cerochroa brachialis</i> Stål	JJG405	AY646301
Atysites		
<i>Diorhabda</i> sp. Weise	CND712	AY243784
^K <i>Diorhabda elongata</i> (Brullé)	SJK712	AY171446
<i>Megaleruca</i> sp. Laboisière	JJG204	AY243780
<i>Megaleruca</i> sp. Laboisière	JJG309	AY243779
<i>Megaleruca</i> sp. Laboisière	JJG320	AY646302
<i>Pyrrhalta maculicollis</i> (Motschulsky)	JJG190	AY243781
<i>Pyrrhalta aenescens</i> (Fairmaire)	JJG187	AY646303
<i>Pyrrhalta</i> sp. Joannis	JJG316	AY243782
Schematizites		
<i>Metrogaleruca</i> sp. Bechyné & Bechyné	JJG134	AY243777
<i>Monoxia debilis</i> LeConte	JJG239	AY243778
<i>Neolachmaea dilatipennis</i> (Jacoby)	JJG323	AY243785
<i>Ophraea</i> sp. Jacoby (or nr.)	JJG131	AY243770
<i>Ophraella notulata</i> (Fabricius)	JJG095	AY243783
<i>Schematiza flavofasciata</i> (Klug)	JJG188	AY243786
^K <i>Schematiza flavofasciata</i> (Klug)	ZSH003	AY171447
Apophyllites (apo)		
<i>Pseudadimonia variolosa</i> (Hope)	JJG312	AY243775
<i>Apophyllia pallipes</i> (Baly)	JJG429	AY646304

Table 5 Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section)	Extract Code ^b	Accession Number
Metacyclini		
New World genera		
<i>Chthoneis</i> sp. Baly	JJG109	AY243764
<i>Chthoneis</i> sp. Baly (nr. <i>marginicollis</i>)	JJG354	AY646305
<i>Chthoneis</i> sp. Baly (nr. <i>iquitoensis</i>)	JJG361	AY646306
<i>Masurius violaceipennis</i> (Jacoby) (or nr.)	JJG116	AY243766
<i>Malachorhinus sericeus</i> Jacoby	JJG129	AY243765
<i>Exora obsoleta</i> (Fabricius)	JJG110	AY243762
<i>Exora obsoleta</i> (Fabricius)	JJG353	AY243763
<i>Exora</i> sp. Chevrolat	JJG340	AY646307
<i>Pyesia</i> sp. Clark	JJG246	AY243767
<i>Zepherina</i> sp. Bechyné (or nr.)	JJG342	AY646308
Old World genus		
<i>Palaeophylia</i> sp. Jacoby (or nr.)	JJG222	AY243768
Hylaspini		
Antiphites		
<i>Pseudeusttetha hirsuta</i>	JJG443	AY646309
<i>Emathea subcaerulea</i>	JJG442	AY646310
Sermylites		
<i>Aplosonyx orientalis</i> (Jacoby)	JJG436	AY646311
<i>Aplosonyx quadriplagiatus</i> (Baly)	JJG173	AY243675
<i>Aplosonyx</i> sp. Chevrolat	JJG427	AY646312
<i>Aplosonyx</i> sp. Chevrolat	JJG412	AY646313
<i>Sermylassa halensis</i> (Linnaeus)	JJG179	AY243676
Hylaspites		
<i>Agelasa nigriceps</i> Motschulsky	JJG319	AY243677
<i>Doryidella</i> sp. Laboissière (or nr.)	JJG425	AY646314
<i>Sphenoraia paviei</i> Laboissière	JJG437	AY646315
Agelastites		
<i>Agelastica coerulea</i> Baly	JJG315	AY243678
^K <i>Agelastica coerulea</i> Baly	SJK701	AY171425
Luperini		
Luperini Chapuis "genus undet."	JJG376	AY646338
Aulacophorina		
Aulacophorites		
<i>Paridea</i> sp. Baly (or nr.)	JJG235	AY243696
<i>Chosnia obesa</i> (Jacoby) (or nr.)	JJG201	AY243697
<i>Sonchia sternalis</i> Fairmaire (or nr.)	JJG210	AY243698
<i>Aulacophora indica</i> (Gmelin)	JJG220	AY243701
^K <i>Aulacophora indica</i> (Gmelin)	SJK711	AY171444
<i>Aulacophora lewisii</i> Baly	JJG158	AY243700
<i>Aulacophora lewisii</i> Baly	JJG228	AY243699
<i>Aulacophora lewisii</i> Baly	JJG127	AY646316
<i>Leptaulaca fissicollis</i> Thomson (or nr.)	JJG234	AY243703
<i>Diacantha fenestrata</i> Chapuis (or nr.)	JJG232	AY243704
Idacanthites		
<i>Prosmidia conifera</i> Fairmaire (or nr.)	JJG212	AY243702
Diabroticina		
Diabroticites		
Diabroticites Chapuis "genus undet."	JJG345	AY646339
<i>Isotes multipunctata</i> (Jacoby)	JJG300	AY243723
<i>Isotes</i> sp. Weise	JJG145	AY243724
<i>Isotes</i> sp. Weise	JJG349	AY243722

Table 5 Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section)	Extract Code ^b	Accession Number
<i>Isotes</i> sp. Weise	JJG351	AY243720
<i>Isotes</i> sp. Weise	JJG363	AY243721
<i>Isotes</i> sp. Weise	JJG372	AY243725
<i>Isotes</i> sp. Weise	JJG373	AY243726
<i>Paranapiacaba tricineta</i> (Say)	JJG322	AY243753
<i>Paranapiacaba</i> sp. Bechyné	JJG094	AY243752
<i>Acalymma vittatum</i> (Fabricius)	JJG413	AY646317
<i>Acalymma fairmairei</i> (Baly)	JJG016	AY243708
<i>Acalymma bivittatum</i> (Fabricius)	JJG297	AY243709
<i>Acalymma blomorum</i> Munroe & R. Smith (or nr.)	JJG229	AY243710
<i>Acalymma trivittatum</i> (Mannerheim)	JJG059	AY243711
<i>Acalymma hirtum</i> (Jacoby)	JJG053	AY243712
<i>Acalymma albidovittatum</i> (Baly)	JJG305	AY243713
<i>Acalymma</i> sp. Barber	JJG359	AY243714
<i>Acalymma</i> sp. Barber	JJG360	AY243715
<i>Acalymma</i> sp. Barber	JJG399	AY646318
<i>Paratriarius subimpressa</i> (Jacoby)	JJG128	AY243727
<i>Paratriarius</i> sp. Schaeffer	JJG147	AY243728
<i>Paratriarius</i> sp. Schaeffer	JJG348	AY243729
<i>Paratriarius</i> sp. Schaeffer	JJG374	AY243730
<i>Amphelasma nigrolineatum</i> (Jacoby)	JJG227	AY243754
<i>Amphelasma sexlineatum</i> (Jacoby)	JJG295	AY243755
<i>Diabrotica balteata</i> LeConte	JJG288	AY243731
<i>Diabrotica biannularis</i> Harold	JJG010	AY243732
<i>Diabrotica decempunctata</i> (Latreille)	JJG299	AY243733
<i>Diabrotica speciosa</i> (Germar)	JJG306	AY646319
<i>Diabrotica speciosa speciosa</i> (Germar)	JJG125	AY271865
<i>Diabrotica virgifera virgifera</i> LeConte	JJG060	AY243734
<i>Diabrotica adelpha</i> Harold	JJG046	AY243735
<i>Diabrotica porracea</i> Harold	JJG292	AY243737
<i>Diabrotica undecimpunctata howardi</i> Barber	JJG370	AY243739
<i>Diabrotica undecimpunctata howardi</i> Barber	JJG223	AY243738
^k <i>Diabrotica undecimpunctata howardi</i> Barber	SJK223	AY171445
<i>Diabrotica tibialis</i> Jacoby	JJG170	AY243746
<i>Diabrotica limitata</i> (Sahlberg)	JJG313	AY243747
<i>Diabrotica l. quindecimpunctata</i> (Germar)	JJG180	AY243736
<i>Diabrotica viridula</i> (Fabricius)	JJG314	AY243748
<i>Diabrotica</i> sp. Chevrolat	JJG335	AY243740
<i>Diabrotica</i> sp. Chevrolat	JJG336	AY243741
<i>Diabrotica</i> sp. Chevrolat	JJG341	AY243742
<i>Diabrotica</i> sp. Chevrolat	JJG356	AY243743
<i>Diabrotica</i> sp. Chevrolat	JJG362	AY243744
<i>Diabrotica</i> sp. Chevrolat	JJG365	AY243745
<i>Gynandrobrotica nigrofasciata</i> (Jacoby)	JJG152	AY243717
<i>Gynandrobrotica lepida</i> (Say)	JJG298	AY243718
<i>Gynandrobrotica</i> sp. Bechyné	JJG358	AY243716
<i>Gynandrobrotica</i> sp. Bechyné	JJG371	AY243719
<i>Gynandrobrotica ventricosa</i> (Jacoby)	JJG135	AY646321
Ceratomytes		
<i>Neobrotica caeruleofasciata</i> Jacoby	JJG117	AY243749
<i>Neobrotica</i> sp. Jacoby	JJG337	AY243750
<i>Neobrotica</i> sp. Jacoby	JJG375	AY243751
<i>Eucerotoma</i> sp. Laboissière	JJG344	AY243756

Table 5 Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section)	Extract Code ^b	Accession Number
<i>Eucerotoma</i> sp. Laboissière	JJG346	AY243759
<i>Eucerotoma</i> sp. Laboissière	JJG347	AY243757
<i>Eucerotoma</i> sp. Laboissière	JJG364	AY243758
<i>Cerotoma arcuata</i> (Olivier)	JJG048	AY243760
<i>Cerotoma</i> sp. Chevrolat	JJG339	AY243761
<i>Cerotoma ruficornis</i> (Olivier)	JJG172	AY646322
<i>Cerotoma facialis</i> Erichson	JJG161	AY646323
Phyllecthrites		
<i>Trichobrotica nymphaea</i> Jacoby	JJG226	AY243706
<i>Phyllecthris gentilis</i> LeConte	JJG366	AY243707
<i>Phyllecthris</i> Dejean "genus undet."	JJG377	AY646324
Trachyscelidites		
<i>Trachyscelida</i> sp. Horn	JJG224	AY243705
Luperina		
Adoxiites		
<i>Medythia suturalis</i> (Motschulsky)	JJG434	AY646325
<i>Medythia suturalis</i> (Motschulsky)	JJG448	AY646326
Scelidites		
<i>Scelolyperus leontii</i> (Crotch)	JJG099	AY243684
<i>Scelolyperus meracus</i> (Say)	JJG257	AY243686
<i>Scelolyperus</i> sp. Crotch	JJG054	AY243685
<i>Lygistus streptophallus</i> Wilcox	JJG367	AY243687
<i>Keithaeus blakeae</i> (White)	JJG414	AY646327
<i>Stenoluperus nipponensis</i> Laboissière	CND717	AY243694
Phyllobroticites		
<i>Phyllobrotica</i> sp. Chevrolat	JJG076	AY243690
^k <i>Phyllobrotica</i> sp. Chevrolat	SJK076	AY171427
<i>Mimastra gracilicornis</i> Jacoby	JJG287	AY243691
<i>Mimastra</i> sp. Baly	JJG430	AY646328
<i>Hoplasoma unicolor</i> Illiger	JJG419	AY646329
Ornithognathites		
<i>Hallirhotius</i> sp. Jacoby	JJG206	AY243689
Exosomites		
<i>Pteleon brevicornis</i> (Jacoby)	JJG415	AY646330
<i>Liroetiella bicolor</i> Kimoto	JJG368	AY646331
<i>Cassena indica</i> (Jacoby)	JJG416	AY646332
Monoleptites		
Monoleptites Chapuis "genus undet."	JJG422	AY646333
Monoleptites Chapuis "genus undet."	JJG431	AY646334
Monoleptites Chapuis "genus undet."	JJG440	AY646335
Monoleptites Chapuis "genus undet."	JJG338	AY646296
<i>Monolepta nigrotibialis</i> Jacoby	JJG044	AY243681
^k <i>Monolepta nigrotibialis</i> Jacoby	SJK044	AY171426
<i>Monolepta</i> sp. Chevrolat	JJG183	AY243682
<i>Monolepta</i> sp. Chevrolat	JJG310	AY243679
<i>Monolepta</i> sp. Chevrolat	JJG317	AY243680
<i>Monolepta</i> sp. Chevrolat	JJG369	AY243683
<i>Metrioidea</i> sp. Fairmaire (or nr.)	JJG301	AY243688
Luperites		
<i>Spilocephalus bipunctatus</i> Allard	JJG205	AY243692
<i>Palpoxena</i> sp. Baly	JJG230	AY243693
<i>Luperus longicornis</i> Fabricius	JJG407	AY646336
Megalognathites		

Table 5 Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section)	Extract Code ^b	Accession Number
<i>Megalognatha</i> sp. Baly	JJG303	AY646337
Unidentified specimens		
Thailand specimen 4	JJG411	AY646340
Thailand specimen 7	JJG417	AY646341
Thailand specimen 8	JJG418	AY646342
Thailand specimen 10	JJG420	AY646343
Thailand specimen 11	JJG421	AY646344
Thailand specimen 13	JJG423	AY646345
Thailand specimen 14	JJG424	AY646346
Thailand specimen 22	JJG432	AY646347
Thailand specimen 25	JJG435	AY646348
Thailand specimen 31	JJG441	AY646349
Thailand specimen 36	JJG446	AY646350
Thailand specimen 37	JJG447	AY646351

^a Taxonomic groupings follow Seeno & Wilcox (1982).

^b DNA extraction codes for all taxa are listed as recorded on all vouchered specimens.

^k Denotes sequence from Kim *et al.* (2003).

Elmer) 377 automated DNA sequencer. Both anti-sense and sense strands were sequenced for all taxa, and edited manually with the aid of Sequence Navigator™ (Applied Biosystems). During editing of each strand, nucleotides that were readable, but showed either irregular spacing between peaks, or had some significant competing background peak, were coded with lower case letters or IUPAC-IUB ambiguity codes. Consensus sequences were exported into Microsoft Word™ for manual alignment.

Multiple sequence alignment

The 28S-D2,D3 sequences were aligned manually according to secondary structure, with the notation following Kjer *et al.* (1994) and Kjer (1995), with slight modifications (see Fig. 2). Alignment initially followed the secondary structural models of Gutell *et al.* (1994), which were obtained from the website <http://www.rna.icmb.utexas.edu>; Cannone

et al. (2002)), and was further modified according to an existing chrysomelid D2 model (Gillespie *et al.*, 2003, 2004) and a trichopteran D3 model (Kjer *et al.*, 2001). Individual sequences, especially hairpin-stem loops, were evaluated in the program *mfold* (version 3.1; <http://bioinfo.math.rpi.edu/~zukerm/>), which folds rRNA based on free energy minimizations (Matthews *et al.*, 1999; Zuker *et al.*, 1999). These free energy-based predictions were used to facilitate the search for potential base-pairing stems, which were confirmed only by the presence of compensatory base changes across all taxa.

Regions in which positional homology assessments were ambiguous across all taxa were defined according to structural criteria as in Kjer (1997), and described as regions of alignment ambiguity (RAA) or regions of slipped-strand compensation (RSC; Levinson & Gutman, 1987; for reviews regarding rRNA sequence alignment see Schultes *et al.*, 1999; Hancock & Vogler, 2000). Briefly, ambiguous regions in which base-pairing was not indentifiable were characterized as RAAs. For ambiguous regions wherein base-pairing was observed (RSCs), compensatory base change evidence was used to confirm structures that were not consistent across the alignment due to the high occurrence of unknown insertion and deletion events (indels). For two ambiguous regions in the alignment caused by the expanding and contracting of hairpin-stem loops, RSCs were further characterized as RECs (regions of expansion and contraction) based on structural evidence used to identify separate non-pairing ambiguous regions of the alignment (terminal bulges). A recent paper addresses the characterization of RAAs, RSCs and RECs with a discussion on phylogenetic methods accommodating these regions (Gillespie, 2004).

Our alignment was entered into the alignment editor AE2 (developed by T. Macke; see Larsen *et al.* 1993) for comparison to established eukaryotic secondary structural models (Gutell & Fox, 1988; Gutell *et al.*, 1990; Gutell *et al.*, 1992b; Gutell *et al.*, 1993; Schnare *et al.*, 1996; Cannone *et al.*, 2002). This process searched for compensating base changes using computer programs developed within the Gutell laboratory (University of Texas at Austin, <http://www.rna.icmb.utexas.edu/>; discussed in Gutell *et al.*, 1985 and Gutell *et al.*, 1992a) and used subsequent information to infer additional secondary structural features. This refined alignment was reanalyzed for positional covaryations and the entire process was repeated until the proposed structures were entirely compatible with the alignment. Secondary structure diagrams were generated interactively with the computer program XRNA (developed by B. Weiser and H. Noller, University of Santa Cruz). Individual secondary structure diagrams are available at <http://www.rna.icmb.utexas.edu/> and <http://hisl.tamu.edu>. Our complete multiple sequence alignment is posted at <http://hisl.tamu.edu>, with specific explanations regarding the rRNA structural alignment. The reader is encouraged to check the homepage of JJG (<http://hisl.tamu.edu>) for continuing updates to the alignment and availability of secondary structure diagrams.

Comparative sequence analysis

The nucleotide frequency data and covarying positions were obtained with the Sun Microsystems Solaris-based program query (Gutell lab, unpublished software). Positional covariation was identified by several methods including mutual information

(Gutell *et al.*, 1992a), a pseudo-phylogenetic event scoring algorithm (Gautheret *et al.*, 1995), and an empirical method (Cannone *et al.*, 2002). This output was filtered to include only mutual best scores, i.e., pairs of positions that share a high covariation score, and examined for nested patterns that could represent helical regions (Goertzen *et al.*, 2003). These patterns included canonical (G:C and A:U) as well as non-canonical (G:U and C:A) base pairings that are adjacent and antiparallel to one another in helical regions. Nucleotide frequency tables for all positions (excluding RAAs, RSCs and RECs) within the putative "stem-loop" regions were prepared to assess the quality and consistency of the predicted base pairing. In general, we accepted only those base pairs that exhibit near-perfect positional covariation in the dataset or invariant nucleotides with the potential to form Watson-Crick pairings within the same helix (Goertzen *et al.*, 2003).

Our alignment was also modified as a NEXUS file to estimate transition/transversion (ts/tv) ratios. In PAUP* (Swofford, 1999), a heuristic parsimony search implementing 100 random sequence additions, saving 100 trees per replicate (all other settings were left as default), generated 500 equally parsimonious trees. These trees were then used to calculate the mean ts/tv ratios in pairing and non-pairing regions across the entire alignment using the "state changes and statistics" option in the chart menu of MacClade 4.0 (Maddison & Maddison, 2000).

CHAPTER III

CHARACTERIZING REGIONS OF AMBIGUOUS ALIGNMENT CAUSED BY THE EXPANSION AND CONTRACTION OF HAIRPIN-STEM LOOPS IN RIBOSOMAL RNA MOLECULES*

Introduction

Typical phylogenetic studies employing ribosomal RNA-encoding DNA (rDNA) as a marker of choice are faced with the difficulty of generating objectively aligned multiple sequences. Unlike most protein-encoding genes, rDNA sequence alignments across even closely related taxa often contain length heterogeneity, resulting from structurally-less conserved regions with characteristically higher numbers of indels, or unknown insertion/deletion events. Regions characterized by numerous adjacent indels present a high level of ambiguity to any multiple sequence alignment, thus complicating the establishment of positional nucleotide homology. Several solutions have been introduced for the treatment of these ambiguously-aligned regions during phylogenetic analysis, ranging from complete character exclusion (Swofford, 1993) to equal weighting of all sequences regardless of differences in ambiguity and/or

* Reprinted from *Molecular Phylogenetics and Evolution*, Vol. **33**, Gillespie, J.J., Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules, pages 936-943, Copyright (2004), with permission from Elsevier.

structural/functional constraint (reviewed in Lee, 2002). To date, no attempts have been made to distinguish different types of ambiguously aligned regions, despite structural (e.g., Page *et al.*, 2002) and evolutionary differences that actually comprise these features of rRNA molecules.

Another level of complexity accompanying ambiguously-aligned regions in rDNA sequences relates to differences in base composition and rates of change across stem and loop regions. Across a multiple sequence alignment, the use of information from secondary structure to partition rRNA molecules into paired and unpaired regions (Kjer, 1995) can facilitate 1. the identification of structural and functional differences within the molecule, 2. the localization of ambiguously-aligned regions within homologous positions, and 3. the appropriate accommodation of base composition and substitution rates that are unique in stems, loops and ambiguously-aligned regions. As I will describe in this paper, the methodology of structural alignment, coupled with the distinction between singled-stranded unalignable regions and regions of slipped-strand compensation, can be used for the refined treatment of ambiguously-aligned regions formed by expanding and contracting hairpin-stem loops of rRNA molecules. This method retains more information from these unalignable regions than previous methods, adding characters to datasets that are analyzed using parsimony as an optimality criterion. The reader is encouraged to refer to Table 6 for the terminology used here for the description of rRNA structure and the alignment of rDNA sequences.

Table 6. Glossary of terms used for alignment and secondary structure of rRNA ^a.

Term	Definition
Helix (stem)	A right-handed double helix composed of a succession of complementary hydrogen-bonded nucleotides between paired strands.
Single strand (loop)	Unpaired nucleotides separating helices.
Hairpin-stem loop	Helix closed distally by a loop of unpaired nucleotides (terminal bulge).
Terminal bulge	Succession of unpaired nucleotides at the end of a hairpin-stem loop.
Lateral bulge	Succession of unpaired nucleotides on one strand of a helix.
Internal bulge	Group of nucleotides from two antiparallel strands unable to form canonical pairs.
Compensatory base change	Subsequent mutation on one strand of a helix to maintain structure following initial mutation of a (CBC) complementary base.
Insertion	A single insertion of a nucleotide relative to the rest of the multiple sequence alignment.
Deletion	A single deletion of a nucleotide relative to the rest of the multiple sequence alignment.
Indel	An ambiguous position within a multiple sequence alignment that cannot be described as an insertion or deletion.
Region of ambiguous alignment (RAA)	Two or more adjacent, non-pairing positions within a sequence wherein positional homology cannot be confidently assigned due to the high occurrence of indels in other sequences.
Region of slipped-strand compensation (RSC)	Region involved in base-pairing wherein positional homology cannot be defended across a multiple sequence alignment; inconsistency in pairing likely due to slipped-strand mispairing.
Region of expansion and contraction (REC)	Variable helical region flanked by conserved basepairs at the 5' and 3' ends, and an unpaired terminal bulge of at least three nucleotides; characteristic of RNA hairpin-stem loops.

^a Modified from Ouvrard et al. (2000).

Structural perspective

The multiple sequence alignment of rDNA is often problematic when the degree of length heterogeneity amongst taxa is high (De Rijk *et al.*, 1995). While the majority of helices in rRNA molecules are structurally conserved across the most divergent of taxa

(Gutell *et al.*, 1994; Gutell, 1996), some helices and non-pairing regions, such as hairpin-stem loops and terminal and lateral bulges, can vary greatly in nucleotide sequence length and base composition even in closely related taxa (e.g., Hillis & Dixon, 1991; Schnare *et al.*, 1996; Gillespie *et al.*, 2004b). This characteristic of rRNA structure, coupled with the fact that pairing and non-pairing regions often accumulate substitutions at different rates (Van de Peer *et al.*, 1993), suggests that evolutionary studies utilizing these molecules for phylogeny reconstruction should benefit from the *a priori* designation of higher order structure to rDNA sequences. For instance, several studies have shown that structural information provides an objective criterion for assigning positional nucleotide homology in difficult-to-align rDNA datasets (e.g., Kjer, 1995; Hickson *et al.*, 1996; Kjer, 1997; Noterdame *et al.*, 1997; Hwang *et al.*, 1998; Lutzoni *et al.*, 2000; Goertzen *et al.*, 2003; Xia *et al.*, 2003). Also, Hickson *et al.* (2000) demonstrated that automated alignment methods fail to align sequences according to their conserved structural motifs, undoubtedly a consequence of these algorithms being based on phenetic sequence distance as opposed to structures that are more conserved than primary nucleotide sequence. Some studies have even shown that alignments based on structural information improve phylogeny estimation (Dixon & Hillis, 1993; Kjer, 1995; Titus & Frost, 1996; Morrison & Ellis, 1997; Uchida *et al.*, 1998; Mugridge *et al.*, 1999; Cunningham *et al.*, 2000; Gonzalez & Labarere, 2000; Hwang & Kim, 2000; Lydeard *et al.*, 2000; Morin, 2000; Xia, 2000; Xia *et al.*, 2003). A recent example of this is the study of Xia *et al.* (2003) in which only structural alignments (and appropriate substitution models based on structure) were able to recover the well-accepted

phylogeny of tetrapods using "analytically-challenging" nuclear SSU rDNA (18S) sequences.

Structural alignment of rDNA sequences involves using published structural models to facilitate the alignment of conserved motifs, with more variable regions aligned by searching for compensatory base changes (CBCs, Table 6) in complementary pairing-regions (Kjer, 1995). This process is performed throughout the entire alignment, leaving unalignable regions delimited by objectively-defined homologous characters (Kjer, 1997; Lutzoni *et al.*, 2000). These unalignable regions, wherein the assignment of positional homology cannot be confidently defended, are initially excluded from phylogenetic analyses (e.g., commented-out in the alignment file). Several studies have suggested ways to retain information from these structurally-defined ambiguous regions (Shapiro & Zhang, 1990; Hofacker *et al.*, 1994; Nedbal *et al.*, 1994; Baldwin *et al.*, 1995; Hibbet *et al.*, 1995; Kjer, 1995; Crandall & Fitzpatrick, 1996; Kretzer *et al.*, 1996; Manos, 1997; Billoud *et al.*, 2000; Collins *et al.*, 2000; Flores-Villela *et al.*, 2000; Lutzoni *et al.*, 2000; Lydeard *et al.*, 2000; Ouvrard *et al.*, 2000; Kjer *et al.*, 2001; Manuel *et al.*, 2003). An example most relevant to this study is that of Lutzoni *et al.* (2000) in which a program was introduced, INAASE, which provides an unequivocal coding method that calculates transformation costs between ambiguously-aligned regions across an alignment without violating positional homology. In this method, unalignable regions, which may be delimited by secondary structure, are initially excluded from homology assignment and coded as single multistate characters defined by the unique combination of nucleotides in each region. These characters are

subsequently assigned costs from step matrices that account for the differential number of changes required to transform each ambiguous region to another. These reweighted characters can then be combined with characters from the unambiguously-aligned sequences under the parsimony optimality criterion for phylogenetic analysis (e.g., combined in PAUP* (Swofford, 1999)).

While INAASE is very practical in the retention of phylogenetic information from ambiguously-aligned regions without simultaneous analysis with the unambiguously-aligned sequences, the approach has several computational limitations, as pointed out by the authors (Lutzoni *et al.*, 2000). One of these restrictions of INAASE is that most phylogenetic programs have a limit on the number of states that can be assigned to a character (i.e., in PAUP* only 32 character states can be defined for each multistate ambiguous region; other common programs allow far fewer). This limitation poses difficulty on retaining information from lengthy unalignable regions, as the character states generated in INAASE quickly become exhausted with a high number of nucleotides in the recoded multistate characters. Thus, methods for simplifying ambiguously-aligned regions to retain as much information as possible using the Lutzoni *et al.* (2000) method are desirable (e.g., Kjer *et al.*, 2001).

RAA/RSC/REC coding

In the alignment of rDNA sequences, ambiguously aligned regions arise from the comparison of less-conserved structures of the rRNA molecules. Regions accumulating insertions and deletions are typically single-stranded motifs, such as lateral and internal

bulges, as well as terminal bulges (loops) formed by the expansion and contraction of hairpin helices (stems). Because the absence of base-pairing can be demonstrated in these regions by a lack of CBC evidence, I suggest that they be characterized as regions of ambiguous alignment (RAA) based solely on the high occurrence of adjacent indels across a multiple sequence alignment (Table 6).

Alternatively, some ambiguously-aligned regions, wherein covariation of bases is observed, arise as a result of inconsistency in base-pairing across positions of an alignment. In these regions, CBC evidence can be used to confirm structures that are not consistent across the alignment due to the high occurrence of indels. These regions, wherein selection on helical formation is less conserved than in other helices, are characterized by likely slipped-strand mutation events arising in DNA replication (Schultes *et al.*, 1999; Hancock & Vogler, 2000). While CBCs are often observable in these regions via slipped-strand compensation, the presence of indels makes the assessment of columnar homology difficult, with positional homology often impossible to assign due to different positions within the alignment hydrogen-bonding to one similar position. Thus, because these ambiguously-aligned regions are discretely different than RAAs, I suggest classifying them as regions of slipped-strand compensation (RSC, Table 6) as a description of the inconsistency in base-pairing that characterizes them.

When comparing very divergent taxa in a multiple sequence alignment, the most complex ambiguously-aligned regions, both in length heterogeneity and base composition, are often those formed by expanding and contracting hairpin-stem loops

(Crease & Taylor, 1998). I suggest that the terminal pairing regions of hairpin-stem loops are distinct from other ambiguously-aligned regions (RAA, RSC) in that, while homology assignment is further complicated by slipped-strand mispairing, secondary structural information across the alignment can be used to isolate the terminal bulge (Table 6) formed by the expanding and contracting hairpin helix. This process requires enough sequence conservation to unambiguously define the boundaries of three elements in these regions: 1. the conserved distal helix, 2. the apical pairing region of the helix that is, in essence, an RSC, and 3. the unpaired terminal bulge, which in rRNA molecules must be at least three nucleotides in length. If high levels of variation in sequence length and base composition occur in the terminal bulge, then this region becomes an RAA by the definition described above. This manual dissection of the terminal bulge requires careful inspection of the distal pairing-regions of the helix, which often accumulate many indels, even between closely related taxa (e.g., Crease & Taylor, 1998). I suggest the term region of expansion and contraction (REC, Table 6) to describe the apical helical component in highly variable hairpin-stem loops of rRNA molecules.

This dissection of rRNA hairpin-stem loops into 2 less-conserved complimentary RSCs (RECs) and 1 often highly-conserved terminal bulge (or RAA if unalignable) is demonstrated in detail here. The reduction of one ambiguous region into three is promising for retaining phylogenetic signal from these regions for their subsequent coding as multi-state characters using the parsimony optimality criterion (i.e., for the program INAASE). Also, separating pairing-ambiguous regions (RECs, RSCs) from

non-pairing ambiguous regions (RAAs) allows for the designation of more appropriate substitution models in these evolutionarily distinct regions of rRNA.

Results and discussion

A predicted secondary structure of the expansion segment D2 from the 28S rRNA is shown in Figure 8. The boxed regions correspond to two motifs of the molecule that terminate in hairpin-stem loops. The alignment generated for the first boxed region ("helix 2" in Gillespie *et al.*, 2004b) from 11 representative chrysomelid beetles is shown in Figure 9 (A-C). The data are shown unaligned (Fig. 9A), aligned according to secondary structure following the convention of Kjer (1995) (Fig. 9B), and further modified according to the RAA/RSC/REC coding described here (Fig. 9C). Indel 3 (I3) from Figure 9B is further subdivided into three ambiguously-aligned regions, with **REC(1)** and **REC(1')** depicting the expansion and contraction of the hairpin-stem loop formed by the homologous **2f** helix (Fig. 9C). Despite the presence of insertions and deletions (indels) in these regions, it is still possible to use CBC evidence to "carve out" the non-pairing terminal bulge. This terminal bulge, **RAA(2)**, is formed by the folding of **REC(1)** and **REC(1')** and represents those nucleotides that are non-pairing across the alignment in this region. In this example, an RSC is formed between the conserved helices **2e** and **2f** (labeled **RSC(1)**). Although no indel events occur in this region, positional homology cannot be assigned due to the inconsistency of columnar hydrogen-bonding, a result probably attributed to slipped-strand mispairing (see Gillespie *et al.*, 2004b for more details on this ambiguously-aligned region).

Figure 9. Demonstration of the subdivision of ambiguously-aligned regions of rDNA sequences formed by the expansion and contraction of hairpin-stem loops. (A-C) The first region ("helix 2") of the 28S expansion segment D2 boxed in Figure 1. (D-F) The second region ("helix 3-1") of the 28S expansion segment D2 boxed in Figure 8. Taxa (and GenBank accession numbers) included in the alignments are: (a) *Lamprosoma* sp. ([AY243651](#)), (b) *Microrhopala vittata* ([AY243650](#)), (c) *Timarcha tenebricosa* ([AY171439](#)), (d) *Chaetocnema costulata* ([AY171431](#)), (e) *Systema* sp. ([AY243665](#)), (f) *Sonchias sternalis* ([AY243698](#)), (g) *Gynandrobrotica ventricosa* ([AY646321](#)), (h) *Chthoneis* sp. ([AY243764](#)), (i) *Caraguata pallida* ([AY243776](#)), (j) *Diorhabda* sp. ([AY243784](#)), (k) *Monocesta* sp. ([AY171433](#)), (l) *Orsodacne atra* ([AY243660](#)), (m) *Chalcophana* sp. ([AY243658](#)), (n) *Chrysolina coerulaus* ([AY171429](#)), (o) *Altica* sp. ([AY171424](#)), (p) *Walterianella bucki* ([AY243673](#)), (q) *Agelastica coerulea* ([AY243678](#)), (r) *Megalognatha* sp. ([AY646337](#)), (s) *Eucerotoma* sp. ([AY243756](#)), (t) *Monolepta nigrotibialis* ([AY243681](#)), (u) *Megaleruca* sp. ([AY646302](#)), and (v) *Neolochmaea dilatipennis* ([AY243785](#)). (A,D) Unaligned sequences. (B,E) Sequences aligned according to secondary structure following the convention of Kjer et al. (1994) and Kjer (1995) with minor alterations described in Gillespie et al. (2004b). (C,F) The modified structural alignment showing the subdivision of ambiguously-aligned regions in distal positions of hairpin-stem loops. Terminal bulges and recoded characters are in bold. Helices with long-range interactions are placed within bars (|) and immediate hairpin-stem loops are placed within double bars (||). All complimentary strands are depicted with a prime (e.g., strand **2e** hydrogen bonds with strand **2e'** to form helix **2e**). Regions of alignment ambiguity (RAA), slipped-strand compensation (RSC) and expansion and contraction (REC) are placed within brackets ([]) and described in the text. Nucleotides within helices involved in hydrogen-bonding are underlined. Note: underscoring of positions involved in hydrogen-bonding in RSCs and RECs do not depict homologous basepairs across the alignment, but portray the distinct structures formed in each taxon. Single insertions (*) and deletions (-) are noted as in Kjer et al. (2001). Positions that can form an expansion of a helix across some but not all taxa are labeled with a hat (^). The sequences are in 5'-3' direction. Missing nucleotides are represented with question marks (?).

a Lamprosoma UGGAGCGUCCGAGUGUGACGGGAUGACGCGCCGCGUUUACGCGUGCGUCUGUCGUCGCCGCGCCUUUCGUUCUUCG

b Microrhopala UUGAGCGGGUUGCGGACGGUCGAGGUACGCUUUCGUCGCCACUUUCGUGUUUUUCGA

c Timarcha AUAAACGAGCGAGUGACGGACGACGUAUGUUUCGCGUACGCUUUUCGUUUUCGUUUAGGU

d Chaetocnema UUGGACGUUCGAGUGACGGACGACGCUUCGCGCGUCGUCUGCGCCGUUUCGUUCAUGCA

e Systena UUGAGCGUUUCGGCGACGAUUGCGUUUUUAUUGUCGUUUUGGCAUUCUUCGUUCUUCGA

f Sonchia UUGAACCGUUUAUACGACGGGAUGGCUUUCGAGUUGUCUGCGUUUUUAUGUUUUUCGA

g Gynandrobrotica UUGGAUGUUUGCGAUGACGGGAUAAACGUUUUCGCGUGUCUGCGUUUCUUAUAUUCUUUGA

h Chthoneis UUGGACGUUUUGGACGGGAUGGUGUUUCGCGACAUUCGUCGCUUAUUGUUUCUUCGA

i Caraguata UUGAACGUUUUUUAUGACGGGAUUAUGUGUGCUUUCGCGUUUACCCGAUCCGCAUAUUAUGUUUUUCGA

j Diorhabda UUGGAACCGUUUGUUAUGACGGGAUUAUGUUUUGUUUCGCGCAUAUUUACAAUAAAUACGCUUUUACGUUUUUUCGA

k Monocesta UUGAAUUGUUUUUGAUUGACGGGAUGACGUGUGCGCUUCGGGCUUCUCUACUGUUUAUCCGCAUAUUAUUCUUCGGA

	2d	2e	I	I	2f	I	2f'	I	2e'	I	2d'	
		*	(1)	(2)		(3)		(4)		(5)		
a	UGG	AG-CGU	[CC]	[GAGUGUG] A	CGGA	[UGACGCGCGCGUUUACGCGUGCGUCGUCG---]	0	UCCG	[CGCCUU]	UCGUU	[-CU]	UCG
b	UUG	AG-CGG	[--]	[UUUCGUG] A	CGGU	[CAGGUUACGCUUCG---]	1	UCCG	[CACCUU]	UCGUU	[UUUU]	CGA
c	AUA	AA-CGA	[--]	[GCAGUG] A	CGGA	[CGACGUACGUUCGCGUCUACGUUU-----]	2	UCUG	[UUUCUU]	UCGUU	[-UA]	GGU
d	UUG	GA-CGU	[--]	[UCGAGUG] A	CGGA	[CGACGUUUCGCGUCG---]	3	UCUG	[CGCCUG]	UCGUU	[CAU]	CGA
e	UUG	AG-CGU	[--]	[UUCGGCG] A	CGAA	[UGACGUUUCGUAUGUCG---]	4	UUUG	[AGUCAU]	UCGUU	[UUU]	CGA
f	UUG	AA-CGU	[--]	[UUAUACG] A	CGGA	[UGGCUUUCGAUGUG-----]	5	UCUG	[CGUUUU]	AUGUU	[UUUU]	CGA
g	UUG	GA-UGU	[--]	[UUGCAUG] A	CGGA	[UAAUGUUUCGGCGUCG---]	6	UUUG	[UUUUUU]	AUAUC	[UUU]	CGA
h	UUG	GA-CGU	[--]	[UUGUUUG] A	CGGA	[UGGUUUUUCGCGACAUUCG-----]	7	UCCG	[CAUCUU]	AUGUU	[UUU]	CGA
i	UUG	AA-CGU	[--]	[UUUUUUG] A	CGGA	[UUAUUGUGUGUCUUCGCGGUUACACG-----]	8	UCCG	[CAUAUU]	AUGUU	[UUUU]	CGA
j	UUG	AA-CGU	[--]	[UUGUAUG] A	CGGA	[UAUUUUUUUGUUUCGGCAUUAUACAAUAA-----]	9	AUCG	[CGUUUU]	ACGUU	[UUUU]	CUA
k	UUG	AA-UGU	[AU]	[UUGUAUG] A	CGGA	[UGACGUUGUGUGUCUUCGCGUCUCUCUAUGUUUAa	a	UCCG	[CAUAUU]	AUAUU	[UUU]	CGA

	2d	2e	RAA	RSC	2f	RSC	RAA	RSC	2f'	RSC	2e'	RAA	2d'
		*	(1)	(1)		(2)	(2)	(2')		(1')		(3)	
a	UGG	AG-CGU	[CC]	[GAGUGUG] A	CGGA	[UGACGCGCCGCG--]	0	[UUUA-]	0	[-CGCGUGCGUCGUCG]	0	UCCG	[CGCCUU]
b	UUG	AG-CGG	[--]	[UUGCGUG] A	CGGU	[CGAGG-----]	1	[UACG-]	1	[------CUUCG]	1	UCCG	[CACCUU]
c	AUA	AA-CGA	[--]	[GCGAGUG] A	CGGA	[CGACUACGU-----]	2	[UCGC-]	2	[------GCGUACGUUU]	2	UCUG	[CUUCUU]
d	UUG	GA-CGU	[--]	[UCGAGUG] A	CGGA	[CGACGU-----]	3	[UCCG-]	2	[------GCGUCG]	3	UCUG	[CGCCGU]
e	UUG	AG-CGU	[--]	[UUGCGCG] A	CGAA	[UGACGU-----]	4	[UUUAU]	3	[------AUGUCG]	4	UUGU	[AGUCAU]
f	UUG	AA-CGU	[--]	[UUUAUACG] A	CGGA	[UGGCU-----]	5	[UUCG-]	4	[------AUGUG]	5	UCUG	[CGUUUU]
g	UUG	GA-UGU	[--]	[UUGCAUG] A	CGGA	[UAACGU-----]	6	[UUCG-]	4	[------GCGUCG]	6	UUCG	[CUUCUU]
h	UUG	GA-CGU	[--]	[UUGUUUG] A	CGGA	[UGGUGUU-----]	7	[UUCG-]	2	[------GACAUCCG]	7	UCUG	[CAUCUU]
i	UUG	AA-CGU	[--]	[UUUUUUG] A	CGGA	[UUUUGUGUGC-----]	8	[UUCG-]	5	[-GCGGUUACCGA]	8	UCCG	[CAUAAU]
j	UUG	AA-CCGU	[--]	[UUUGUUG] A	CGGA	[UAUUUUUGUGU-----]	9	[UUCG-]	4	[-GCAUUAUUACAAUA]	9	AUCG	[CGUUUU]
k	UUG	AA-UGU	[AU]	[UUUGUUG] A	CGGA	[UGACGUGUGUGCGC]	a	[UUCG-]	4	[GCUUCUCAUGUUUA]	a	UCCG	[CAUAAU]

D

l	Orsodacne	UACGGCCCCCGGUAAGCCCGUCCGGGGUAAACGCUUCGCGGCGUCCCGGGCGGACCGGGCGGUUCCC
m	Chalcophana	UAGACGUUUGCGGUGGAGCACGACGGACGUUUCACGACGUUCGUACGUACCCGUAAAGUCC
n	Chrysolina	UAGGCCCCGAGGUGGAGCCCCAGUGAACGUUUCGCGUUGGACCCUCCGUUCCC
o	Altica	UAGACUUGGGGUGGAGCCCCAGUGGCUUUUUGUCGCGUGGACCCUCCAUGUCC
p	Walterianella	UAAAAUCGGAGUGGUGCCCACGAGUCGUUUACGCGUCACGUGGACCUUUUAUGUAUAUCC
q	Agelastica	UAGACUCCGGGCGGAGCCCCGUGCGGUUAUUUAUUUUAAUUCGUGCGGACCCUCCAUGUCC
r	Megalognatha	UAGGUUCGAGGUGGAGCCCCGUGUAUUUUAAUGUAUUGCGUGGACCCUCCUUAUCCC
s	Eucerotoma	UAGACUUGAGAUGGAGCCCCGCUAAUAUUACGCGGACACUCGAUGUCC
t	Monolepta	UAGGGUUCGAGGUGGAGCCCCAGUAAUUUCCGAUUGCGUGGACCCUUAUGUCC
u	Megaleruca	UAGGCUUCGUGGUGGAGCCAUUGUGAUCGUAUUUAUUGAUGGACCCUCCAUGUCC
v	Neolochmaea	UAGACUUGAGAUCGAAGACCACGUGAUAUUAAUGAAAAUUAAUGAUGCGUGGACUCUCGAUGUCC

E

	3d	*	3e	*	3f	^^	I (1)	^^	3f'	3e'	I (2)	3d'
l	UAC	GGC	-	C-CGCCG	GUU-AGC	CCGUCCG	GG [GUAAACGCUUCGCGGCGU]	0 CC	CGGGCGG	AC	CGGC	CC
m	UAA	GAC	G	UAUGCGG	UGG-AGC	ACGCACG	GA [CGUUUCACGACG-----]	1 UU	CGUACGU	AC	CCGUAA	CC
n	UAA	GGC	-	C-CGAGG	UGG-AGC	CCACGUG	AA [CGUUUCGCG-----]	2 UU	UGCGUGG	AC	CCUCGG	CC
o	UAA	GAC	-	U-UGGGG	UGG-AGC	CCACAUG	GC [UUUU-----]	3 GU	CGCGUGG	AC	CCUCGA	CC
p	UAA	AAA	-	U-CGGAG	UGG-UGC	CCACGAG	UC [GUUUACGC-----]	4 GU	CACGUGG	AC	CUUUUA	CC
q	UAA	GAC	-	U-CGGGG	CGG-AGC	CCGUGCG	GU [UAUUUAUUUUUA-----]	5 AU	CGUGCGG	AC	CCUCGA	CC
r	UAA	GGU	-	U-CGAGG	UGG-AGC	CCGCGUG	AU [UUUUAUGU-----]	6 AU	UGCGUGG	AC	CCUCGU	CC
s	UAA	GAC	-	U-UGAGA	UGG-AGC	CCGCGUA	-- [AUUU-----]	7 --	UACGCGG	AC	ACUCGA	CC
t	UAG	GGU	-	U-CGAGG	UGG-AGC	CCACGUA	AU [UUUC-----]	8 AU	UGCGUGG	AC	CCUCGA	CC
u	UAA	GGC	-	U-CGUGG	UGG-AGA	CCAUGUG	AU [CGUCAUUGAUUG-----]	9 AU	UAUGUGG	AC	CCUCGA	CC
v	UAA	GAC	-	U-CGAGA	UCGAAGA	CCACGUG	AU [AUUAAUGAAAAUUAAUG-]	a AU	UGCGUGG	AC	UCUCGA	CC

F

	3d	*	3e	*	3f	^^	REC (1)	RAA (1)	REC (1')	^^	3f'	3e'	RAA (2)	3d'
l	UAC	GGC	-	C-CGCCG	GUU-AGC	CCGUCCG	GG [GUAAACGCG]	0 [UUC--]	0 [GCGGCGU]	0 CC	CGGGCGG	AC	CGGC	CC
m	UAA	GAC	G	UAUGCGG	UGG-AGC	ACGCACG	GA [CGUU-----]	1 [UAC-]	1 [---GACG]	1 UU	CGUACGU	AC	CCGUAA	CC
n	UAA	GGC	-	C-CGAGG	UGG-AGC	CCACGUG	AA [CGU-----]	2 [UUCG-]	2 [---GCG]	2 UU	UGCGUGG	AC	CCUCGG	CC
o	UAA	GAC	-	U-UGGGG	UGG-AGC	CCACAUG	GC [-----]	3 [UUUU-]	3 [-----]	3 GU	CGCGUGG	AC	CCUCGA	CC
p	UAA	AAA	-	U-CGGAG	UGG-UGC	CCACGAG	UC [GU-----]	4 [UUA-]	4 [---G-]	4 GU	CACGUGG	AC	CUUUUA	CC
q	UAA	GAC	-	U-CGGGG	CGG-AGC	CCGUGCG	GU [UAUU-----]	5 [UUA-]	5 [---UUUU]	5 AU	CGUGCGG	AC	CCUCGA	CC
r	UAA	GGU	-	U-CGAGG	UGG-AGC	CCGCGUG	AU [UU-----]	6 [UAAU-]	6 [---GU]	6 AU	UGCGUGG	AC	CCUCGU	CC
s	UAA	GAC	-	U-UGAGA	UGG-AGC	CCGCGUA	-- [-----]	3 [AUUU-]	7 [-----]	3 --	UACGCGG	AC	ACUCGA	CC
t	UAG	GGU	-	U-CGAGG	UGG-AGC	CCACGUA	AU [-----]	3 [UUCG-]	2 [-----]	3 AU	UGCGUGG	AC	CCUCGA	CC
u	UAA	GGC	-	U-CGUGG	UGG-AGA	CCAUGUG	AU [CGUC-----]	7 [AUU-]	8 [---GAUUG]	7 AU	UAUGUGG	AC	CCUCGA	CC
v	UAA	GAC	-	U-CGAGA	UCGAAGA	CCACGUG	AU [AUUAAU--]	8 [GAAA-]	9 [AUUAAUG]	8 AU	UGCGUGG	AC	UCUCGA	CC

Figure 9 Continued.

The second boxed motif in Figure 8 ("helix 3-1" in Gillespie *et al.*, 2004b) is shown in a multiple sequence alignment of 11 representative chrysomelids in Figure 9 (D-F). The terminal bulge in this example is more difficult to display due to the absence of any hairpin-stem loop expansion past the homologous **3f** helix in several taxa (O, S and T). Furthermore, helix **3f** can be extended by 2 internal positions with the potential for base-pairing in most sampled taxa; however, one taxon (S) cannot form these pairings because at least 3 nucleotides are needed to form a hairpin-stem loop structure in RNA molecules. Thus, while positional homology can be defended with structural evidence in these 2 potential pairing-bases, they are not included within the conserved **3f** helix because they do not represent a consensus, or homologous, structure across the alignment.

The initial exclusion of nucleotide positions from phylogenetic analysis is imperative if the assignment of positional homology cannot be confidently established across a multiple sequence alignment (Kjer, 1995). Because ambiguously-aligned regions often contain valuable phylogenetic signal (Lee, 2001, others therein), it is desirable to retain information from them, albeit without violating positional homology. The division of ambiguously-aligned regions in expanding and contracting hairpin stem-loops into three smaller regions (2 RECs plus 1 RAA or alignable region) based on secondary structure provides a more efficient means for retrieving information from these regions than previous methods that also use structural criteria to delimit unalignable positions in rDNA sequence alignments (Kjer, 1995, 1997; Lutzoni *et al.*, 2000). This is primarily because more information can be obtained from three separate

ambiguously-aligned regions than one large one, either when coding with (Lutzoni *et al.*, 2000) or without (Kjer *et al.*, 2001) assigned transformation costs (Fig. 9C, F). Plus, when taxon sampling is high and sequence lengths increase in heterogeneity, subdividing large ambiguously-aligned regions into smaller components provides 1. a means for comparing structurally-similar nucleotides in fragment level alignment methods (i.e., INAASE), 2. fewer character state transformations between taxa, with less potential to exceed the number of allotted states in a given phylogenetic software, and 3. improvements to existing global structural models for the various rRNA molecules on public databases (e.g., the Comparative RNA Website <http://www.rna.icmb.utexas.edu>). Given that hairpin-stem loops occur frequently in rRNA molecules (e.g., there are 18 hairpin-stem loops in mammalian SSU mitochondrial rRNA), the method described here for retaining information from these often unalignable sequence regions should add support to branches in generated phylogenies using parsimony as an optimality criterion.

Another advantage to the characterization of RSCs and RECs with CBC evidence is that these ambiguously-aligned regions become separated from other unalignable regions that are single-stranded (RAAs). This is ideal because it is known that substitution rates in pairing and non-pairing regions can be quite different (Van de Peer *et al.*, 1993), that U \leftrightarrow C transitions are elevated in pairing regions (Marshall, 1993), and that an overall transition bias occurs in helices (e.g., Rousset *et al.*, 1991; Kraus *et al.*, 1992; Vawter & Brown, 1993; Gatesy *et al.*, 1994; Nedbal *et al.*, 1994; Douzery & Catzeflis, 1995; Springer *et al.*, 1995; Springer & Douzery, 1996) as a means of

repairing deleterious mutations that disrupt base-pairing (Kimura, 1986). Thus, in INAASE, ts/tv ratios more appropriate for pairing-regions can be assigned to RECs and RSCs, with an equal ts/tv ratio applied to non-pairing RAAs. This method of assigning different substitution costs to pairing and non-pairing ambiguous regions further supports Lee's (2001) suggestion that analyzing ambiguously-aligned regions separately from non-ambiguous positions is more appropriate than optimization alignment methods (Mitchison, 1999; Wheeler, 1999) that assign uniform substitution costs across all positions in rDNA sequences. Also, coding these structurally-defined regions separately in INAASE does not require a gap extension penalty, for which in highly length variable hairpin-stem loops, no single set of costs are likely universally optimal (see Petersen *et al.*, 2004).

Some will argue that the two complementary RECs formed by this method are not independent characters, following the original suggestion of Wheeler & Honeycutt (1988) that pairing-regions of rRNA are non-independent characters. These characters, however, are no less independent than the other helices throughout the rRNA molecule, which often account for over 80% of the data (Higgs, 2000). Given that more attention is being paid to addressing the issue of stem interdependence in rRNA molecules (Savill *et al.*, 2000 and others therein; Jow *et al.*, 2002; Hudelot *et al.*, 2003), the next step is to accommodate the non-independence of complementary RECs within a maximum likelihood approach. While an acceptable model for ambiguously-aligned regions has not been formulated, most likely due to the difficulty associated with modeling indels, compositional information within RECs and terminal bulges may be modeled to

determine the directionality of expansion and contraction in an rRNA hairpin-stem loop. For example, taxon S in Figure 3C has the potential to form two additional internal base-pairings to helix **3f**; however, these bonds cannot form thermodynamically because at least 3 nucleotides must form a stable terminal bulge. Thus, this region of the D2 for taxon S is most likely undergoing contraction of the hairpin-stem loop. There are many examples in the expanded alignment from Gillespie *et al.* (2004b; <http://hisl.tamu.edu>) that show evidence for hairpin-stem loop expansion and contraction. Modeling base frequencies and substitution patterns in these regions could prove highly beneficial for understanding the evolution of hairpin-stem loop structures across a range of taxa and for providing yet more phylogenetic information from these valuable character sources in rDNA sequences.

Experimental procedures

The regions of the 28S-D2 expansion segment analyzed here are from three recent phylogenetic studies (Gillespie *et al.*, 2003, Gillespie *et al.*, 2004a; Kim *et al.*, 2003). The alignment method described here is presented in a larger study that provides a refined structural model for the D2-D3 expansion segments and flanking core elements of the 28S rRNA from chrysomelid beetles (Gillespie *et al.*, 2004b). GenBank accession numbers for the sampled taxa are provided in the legends of Figure 2.

Structural alignments followed the conventions of Kjer *et al.* (1994) and Kjer (1995) with slight modifications to the original notation (Gillespie *et al.*, 2004b). Ambiguously-aligned regions within hairpin-stem loops were searched for CBC

evidence across taxa, allowing for at least three unpaired nucleotides to form the terminal bulge. Regions involved in hydrogen-bonding, but wherein positional homology was indefensible, were bracketed and labeled as RSCs or RECs. If alignment of the terminal bulges was ambiguous across all sequences, they too were bracketed and labeled as RAAs. The complete alignment of all 249 chrysomelid taxa from Gillespie *et al.* (2004b) is available with full explanations and related citations at <http://hisl.tamu.edu> and the Comparative RNA Website <http://www.rna.icmb.utexas.edu> (Cannone *et al.*, 2002).

CHAPTER IV

INCORPORATING MAXIMUM LIKELIHOOD MODELS FOR HELICAL AND NON-PAIRING REGIONS OF RIBOSOMAL RNA IN PHYLOGENETIC ANALYSIS: AN EXAMPLE FROM THE 28S LSU rRNA D2 AND D3 EXPANSION SEGMENTS OF ROOTWORMS AND RELATED LEAF BEETLES (COLEOPTERA: CHRYSOMELIDAE; GALERUCINAE)

Overview

Standard models of DNA substitution are not appropriate for analyzing ribosomal-encoding DNA (rRNA) sequences in phylogenetic analysis because of the non-independence of pairing-nucleotides maintaining higher order structure in these molecules. Although many models of RNA sequence evolution have been proposed, a recent study demonstrated that the most general time reversible models outperform models that assume base-pair symmetry and a zero rate of double substitutions across helices (Savill *et al.*, 2001). These maximum likelihood models have now been incorporated into several phylogenetic programs, allowing for the simultaneous analysis of pairing and non-pairing regions of RNA sequences and the inclusion of RNA and DNA models. In this chapter I use one of these programs, PHASE ver. 1.1 (Jow *et al.*, 2002), which includes six RNA models as well as standard DNA maximum likelihood models, to analyze a dataset of 231 structurally-aligned partial 28S rRNA sequences from chrysomelid beetles (Insecta: Coleoptera). Two models each are evaluated for the

three classes of RNA substitution models: 1) six-state models that only consider Watson-Crick pairs and GU UG intermediates; 2) seven-state models that include a mismatch parameter for all non-Watson-Crick base-pairs; and 3) 16-state models that parameterize all non-canonical basepairs separately. I evaluate these models with the use of maximum likelihood optimization criteria and a Bayesian analysis. Patterns of convergence of all model parameters throughout sampling via Markov Chain Monte Carlo simulation are reported. Separate analyses, starting from different seeds, are run for three and six million generations to determine the time until stationarity is reached for all model parameters, as well as model likelihoods, tree priors, and sampled tree lengths. Using the Akaike information criterion and the likelihood ratio test I conclude that models with more assumptions are not statistically superior than models that relax these assumptions, unless enough sampling iterations are performed such that these additional parameters reach convergence. Even under six million generations some model parameters do not reach stationarity, as evident by the estimated sample size required for an efficient mean sampling of the parameter. This implies that caution should be used in implementing highly-parameterized models of RNA substitution, and that sufficient generations should be performed in attempts to reach convergence in model statistics as sampled from their posterior probability distribution. Additionally, regarding datasets of this nature, a typical size for many phylogenetic studies, 16-state models are computationally intractable due to either their high number of parameters or the nature of the parameters that model changes to and from non-canonical pairs, or both.

By ranking the frequency of substitution classes for all models, and evaluating the differences between models of the same class, I identify an asymmetrical rate of double transitions for both six- and seven-state models that has previously been undetected in studies on RNA evolution. Importantly, models that parameterize all four classes of double transitions will not adequately estimate frequencies resulting from this substitution asymmetry. Finally, I take this, as well as the results of all model comparisons, into account as I report on the relative utility of these RNA maximum likelihood models for practical phylogenetic investigations.

Introduction

Ribosomal RNA molecules form secondary and tertiary structures that are highly conserved across divergent organisms, a consequence of the need for preservation of ribosomal function in cellular protein synthesis (Dahlberg, 1989; Wool *et al.*, 1990; Noller, 1991). Higher-order structure in rRNA is obtained primarily through base-pair interactions within the individual RNA molecule (Fresco *et al.*, 1960). Hydrogen-bonding occurs between canonical base-pairs (AU, GC), non-canonical stable (GU) and unstable (AC) intermediates, as well as uncommon GA and AA pairings (Elgavish *et al.*, 2001), to form contiguous, antiparallel structural elements (helices). Other less-frequently occurring base-pairs, as well as other secondary structural elements and tertiary interactions are reviewed in Gutell *et al.* (1994; 2002), and together with conserved secondary structural helices, form the universally-conserved core ribosome that is comparable across all domains of life (Woese *et al.*, 1990a; Winker & Woese,

1991). This organismal universality of the ribosome, coupled with other characteristics such as high copy number of rDNA cistrons per cell and relative ease for primer design in conserved RNA regions, accounts for the commonality of rDNA sequences as markers for phylogeny reconstruction across virtually any lineage of life (Hillis & Dixon, 1991).

Despite a lack of conservation in primary structure, many of the secondary structural elements in rRNA molecules are conserved across all domains of life (Gutell, 1996). Given this, the elucidation of patterns of nucleotide substitution that are characteristic of different rRNA structural motifs (i.e., “stems and loops”) is possible. Wheeler & Honeycutt (1988) identified a directed substitution rate within helices of the 5S rRNA of animals and plants that deviates patterns expected from a neutral model of molecular evolution (Ohta 1973; Kimura 1983). This slightly-deleterious mode of sequence evolution in rRNA, in which non-canonical base-pairings, or bulges, are replaced by compensatory base changes or reversals to the original state, has been identified in subsequent studies (e.g, Rousset *et al.*, 1991; Kraus *et al.*, 1992; Gatesy *et al.*, 1994; Vawter & Brown, 1993; Douzery & Catzeflis, 1995; Springer *et al.*, 1995; Springer & Douzery, 1996) and appears to be the mechanism orchestrating structural conservation in rRNAs.

Paramount to the findings of Wheeler & Honeycutt (1988) was not only the identification of two different selective constraints within the same molecule (pairing versus non-pairing regions), but also the realization that nucleotides within pairing-regions in rRNA datasets are not independent characters. This poses an added difficulty

when treating helices in phylogenetic analysis, as opposed to unpaired nucleotides wherein interdependence with other positions is not easily demonstrated. Wheeler and Honeycutt (1988) suggested separate parsimony analysis of pairing (stems) or non-pairing (loops) regions but not both in simultaneous analysis (. Some workers have implemented a stem-loop-weighting approach to accommodate the non-independence of pairing-regions (Wheeler & Honeycutt, 1988; Smith, 1989; Dixon & Hillis, 1993). However, downweighting stems on the basis of their non-independence will also down-weight positions that are hypervariable, and often non-pairing, thus inaccurately representing the information contained within pairing-regions. Up-weighting compensatory mutations within pairing regions has justification (Ouvrard *et al.*, 2000), particularly if rare substitutions define major clades; however, discerning which characters to weight within an alignment can be puzzling if the ancestral pairing cannot be immediately identified (i.e., before analysis). Finally, assumptions of certain branch support measures such as the bootstrap (Felsenstein, 1985) and the decay index (Bremer, 1988; Donaghue *et al.*, 1992) are violated by the non-independence of rRNA pairing-regions. For all of these reasons a parsimony approach may not adequately accommodate rRNA data. Similarly, standard likelihood models of DNA substitution, which are all based on a 4x4 rate matrix, are also deficient for phylogeny estimation using rRNA due to their failure in accounting for correlated bases forming helices.

Attempts to provide adequate models that account for the evolution of pairing regions in rRNA molecules have been implemented in the last decade. These studies have centered on establishing a substitution matrix that accommodates the non-

independence of helical regions. Unlike the typical 4x4 substitution matrix used for modeling DNA evolution, a matrix modeling rRNA evolution consists of all possible substitutions within a pairing region. Hence, a 16x16 matrix is used to model pairing-regions, with the most general time reversible (GTR; Li & Gu, 1996; Waddell & Steel, 1997) model allowing for 134 free parameters (Fig. 10). Due to the impracticality (Savill *et al.*, 2001) of a GTR 16x16 model, simplifications have been proposed (Table 7). Schöniger & von Haeseler (1994) defined rates in a 16x16 matrix as $r_{ij} = \pi_j$ if states i and j differ by a single substitution, and $r_{ij} = 0$ in the event of double substitutions. Hence, the model has 15 free parameters. Muse (1995) proposed 3 models that simplified the 16x16 substitution matrix even further. The HKY model (after Hasegawa *et al.*, 1985) has 5 free parameters, allows for differential base frequencies across all sites, and distinguishes between substitution frequencies. The second model of Muse (1995) is nearly identical to the HKY model, except that it treats GU and UG as pairings rather than mismatches, and thus more adequately models helices in which a prevalence for stable intermediates is present. The third model of Muse (1995) is highly simplified with only 1 free parameter accounting for 2 rates representing stable base-pairs and mismatches.

Recently, Savill *et al.*, (2001) suggested three additional models with a 16x16 substitution matrix. The first model, 16A of Savill *et al.* (2001), has 19 free parameters, with substitutions limited to five rate classes: single transitions, double transitions, double transversions, all changes to and from non-canonical pairs, and single

		← MM →																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
		AU	GU	GC	UA	UG	CG	AA	AG	AC	GA	GG	CA	CC	CU	UC	UU	
1	AU	*	$\pi_{GU}\alpha_1$	$\pi_{GC}\alpha_2$	$\pi_{UA}\alpha_3$	$\pi_{UG}\alpha_4$	$\pi_{CG}\alpha_5$	$\pi_{AA}\alpha_6$	$\pi_{AG}\alpha_7$	$\pi_{AC}\alpha_8$	$\pi_{GA}\alpha_9$	$\pi_{GG}\alpha_{10}$	$\pi_{CA}\alpha_{11}$	$\pi_{CC}\alpha_{12}$	$\pi_{CU}\alpha_{13}$	$\pi_{UC}\alpha_{14}$	$\pi_{UU}\alpha_{15}$	
2	GU	$\pi_{AU}\alpha_1$	*	$\pi_{GC}\alpha_{16}$	$\pi_{UA}\alpha_{17}$	$\pi_{UG}\alpha_{18}$	$\pi_{CG}\alpha_{19}$	$\pi_{AA}\alpha_{20}$	$\pi_{AG}\alpha_{21}$	$\pi_{AC}\alpha_{22}$	$\pi_{GA}\alpha_{23}$	$\pi_{GG}\alpha_{24}$	$\pi_{CA}\alpha_{25}$	$\pi_{CC}\alpha_{26}$	$\pi_{CU}\alpha_{27}$	$\pi_{UC}\alpha_{28}$	$\pi_{UU}\alpha_{29}$	
3	GC	$\pi_{AU}\alpha_2$	$\pi_{GU}\alpha_{16}$	*	$\pi_{UA}\alpha_{30}$	$\pi_{UG}\alpha_{31}$	$\pi_{CG}\alpha_{32}$	$\pi_{AA}\alpha_{33}$	$\pi_{AG}\alpha_{34}$	$\pi_{AC}\alpha_{35}$	$\pi_{GA}\alpha_{36}$	$\pi_{GG}\alpha_{37}$	$\pi_{CA}\alpha_{38}$	$\pi_{CC}\alpha_{39}$	$\pi_{CU}\alpha_{40}$	$\pi_{UC}\alpha_{41}$	$\pi_{UU}\alpha_{42}$	
4	UA	$\pi_{AU}\alpha_3$	$\pi_{GU}\alpha_{17}$	$\pi_{GC}\alpha_{30}$	*	$\pi_{UG}\alpha_{43}$	$\pi_{CG}\alpha_{44}$	$\pi_{AA}\alpha_{45}$	$\pi_{AG}\alpha_{46}$	$\pi_{AC}\alpha_{47}$	$\pi_{GA}\alpha_{48}$	$\pi_{GG}\alpha_{49}$	$\pi_{CA}\alpha_{50}$	$\pi_{CC}\alpha_{51}$	$\pi_{CU}\alpha_{52}$	$\pi_{UC}\alpha_{53}$	$\pi_{UU}\alpha_{54}$	
5	UG	$\pi_{AU}\alpha_4$	$\pi_{GU}\alpha_{18}$	$\pi_{GC}\alpha_{31}$	$\pi_{UA}\alpha_{43}$	*	$\pi_{CG}\alpha_{55}$	$\pi_{AA}\alpha_{56}$	$\pi_{AG}\alpha_{57}$	$\pi_{AC}\alpha_{58}$	$\pi_{GA}\alpha_{59}$	$\pi_{GG}\alpha_{60}$	$\pi_{CA}\alpha_{61}$	$\pi_{CC}\alpha_{62}$	$\pi_{CU}\alpha_{63}$	$\pi_{UC}\alpha_{64}$	$\pi_{UU}\alpha_{65}$	
6	CG	$\pi_{AU}\alpha_5$	$\pi_{GU}\alpha_{19}$	$\pi_{GC}\alpha_{32}$	$\pi_{UA}\alpha_{44}$	$\pi_{UG}\alpha_{55}$	*	$\pi_{AA}\alpha_{66}$	$\pi_{AG}\alpha_{67}$	$\pi_{AC}\alpha_{68}$	$\pi_{GA}\alpha_{69}$	$\pi_{GG}\alpha_{70}$	$\pi_{CA}\alpha_{71}$	$\pi_{CC}\alpha_{72}$	$\pi_{CU}\alpha_{73}$	$\pi_{UC}\alpha_{74}$	$\pi_{UU}\alpha_{75}$	
MM	7	AA	$\pi_{AU}\alpha_6$	$\pi_{GU}\alpha_{20}$	$\pi_{GC}\alpha_{33}$	$\pi_{UA}\alpha_{45}$	$\pi_{UG}\alpha_{56}$	$\pi_{CG}\alpha_{66}$	*	$\pi_{AG}\alpha_{76}$	$\pi_{AC}\alpha_{77}$	$\pi_{GA}\alpha_{78}$	$\pi_{GG}\alpha_{79}$	$\pi_{CA}\alpha_{80}$	$\pi_{CC}\alpha_{81}$	$\pi_{CU}\alpha_{82}$	$\pi_{UC}\alpha_{83}$	$\pi_{UU}\alpha_{84}$
	8	AG	$\pi_{AU}\alpha_7$	$\pi_{GU}\alpha_{21}$	$\pi_{GC}\alpha_{34}$	$\pi_{UA}\alpha_{46}$	$\pi_{UG}\alpha_{57}$	$\pi_{CG}\alpha_{67}$	$\pi_{AA}\alpha_{76}$	*	$\pi_{AC}\alpha_{85}$	$\pi_{GA}\alpha_{86}$	$\pi_{GG}\alpha_{87}$	$\pi_{CA}\alpha_{88}$	$\pi_{CC}\alpha_{89}$	$\pi_{CU}\alpha_{90}$	$\pi_{UC}\alpha_{91}$	$\pi_{UU}\alpha_{92}$
	9	AC	$\pi_{AU}\alpha_8$	$\pi_{GU}\alpha_{22}$	$\pi_{GC}\alpha_{35}$	$\pi_{UA}\alpha_{47}$	$\pi_{UG}\alpha_{58}$	$\pi_{CG}\alpha_{68}$	$\pi_{AA}\alpha_{77}$	$\pi_{AG}\alpha_{85}$	*	$\pi_{GA}\alpha_{93}$	$\pi_{GG}\alpha_{94}$	$\pi_{CA}\alpha_{95}$	$\pi_{CC}\alpha_{96}$	$\pi_{CU}\alpha_{97}$	$\pi_{UC}\alpha_{98}$	$\pi_{UU}\alpha_{99}$
	10	GA	$\pi_{AU}\alpha_9$	$\pi_{GU}\alpha_{23}$	$\pi_{GC}\alpha_{36}$	$\pi_{UA}\alpha_{48}$	$\pi_{UG}\alpha_{59}$	$\pi_{CG}\alpha_{69}$	$\pi_{AA}\alpha_{78}$	$\pi_{AG}\alpha_{86}$	$\pi_{AC}\alpha_{93}$	*	$\pi_{GG}\alpha_{100}$	$\pi_{CA}\alpha_{101}$	$\pi_{CC}\alpha_{102}$	$\pi_{CU}\alpha_{103}$	$\pi_{UC}\alpha_{104}$	$\pi_{UU}\alpha_{105}$
	11	GG	$\pi_{AU}\alpha_{10}$	$\pi_{GU}\alpha_{24}$	$\pi_{GC}\alpha_{37}$	$\pi_{UA}\alpha_{49}$	$\pi_{UG}\alpha_{60}$	$\pi_{CG}\alpha_{70}$	$\pi_{AA}\alpha_{79}$	$\pi_{AG}\alpha_{87}$	$\pi_{AC}\alpha_{94}$	$\pi_{GA}\alpha_{100}$	*	$\pi_{CA}\alpha_{106}$	$\pi_{CC}\alpha_{107}$	$\pi_{CU}\alpha_{108}$	$\pi_{UC}\alpha_{109}$	$\pi_{UU}\alpha_{110}$
	12	CA	$\pi_{AU}\alpha_{11}$	$\pi_{GU}\alpha_{25}$	$\pi_{GC}\alpha_{38}$	$\pi_{UA}\alpha_{50}$	$\pi_{UG}\alpha_{61}$	$\pi_{CG}\alpha_{71}$	$\pi_{AA}\alpha_{80}$	$\pi_{AG}\alpha_{88}$	$\pi_{AC}\alpha_{95}$	$\pi_{GA}\alpha_{101}$	$\pi_{GG}\alpha_{106}$	*	$\pi_{CC}\alpha_{111}$	$\pi_{CU}\alpha_{112}$	$\pi_{UC}\alpha_{113}$	$\pi_{UU}\alpha_{114}$
	13	CC	$\pi_{AU}\alpha_{12}$	$\pi_{GU}\alpha_{26}$	$\pi_{GC}\alpha_{39}$	$\pi_{UA}\alpha_{51}$	$\pi_{UG}\alpha_{62}$	$\pi_{CG}\alpha_{72}$	$\pi_{AA}\alpha_{81}$	$\pi_{AG}\alpha_{89}$	$\pi_{AC}\alpha_{96}$	$\pi_{GA}\alpha_{102}$	$\pi_{GG}\alpha_{107}$	$\pi_{CA}\alpha_{111}$	*	$\pi_{CU}\alpha_{115}$	$\pi_{UC}\alpha_{116}$	$\pi_{UU}\alpha_{117}$
	14	CU	$\pi_{AU}\alpha_{13}$	$\pi_{GU}\alpha_{27}$	$\pi_{GC}\alpha_{40}$	$\pi_{UA}\alpha_{52}$	$\pi_{UG}\alpha_{63}$	$\pi_{CG}\alpha_{73}$	$\pi_{AA}\alpha_{82}$	$\pi_{AG}\alpha_{90}$	$\pi_{AC}\alpha_{97}$	$\pi_{GA}\alpha_{103}$	$\pi_{GG}\alpha_{108}$	$\pi_{CA}\alpha_{112}$	$\pi_{CC}\alpha_{115}$	*	$\pi_{UC}\alpha_{118}$	$\pi_{UU}\alpha_{119}$
	15	UC	$\pi_{AU}\alpha_{14}$	$\pi_{GU}\alpha_{28}$	$\pi_{GC}\alpha_{41}$	$\pi_{UA}\alpha_{53}$	$\pi_{UG}\alpha_{64}$	$\pi_{CG}\alpha_{74}$	$\pi_{AA}\alpha_{83}$	$\pi_{AG}\alpha_{91}$	$\pi_{AC}\alpha_{98}$	$\pi_{GA}\alpha_{104}$	$\pi_{GG}\alpha_{109}$	$\pi_{CA}\alpha_{113}$	$\pi_{CC}\alpha_{116}$	$\pi_{CU}\alpha_{118}$	*	$\pi_{UU}\alpha_{120}$
		16	UU	$\pi_{AU}\alpha_{15}$	$\pi_{GU}\alpha_{29}$	$\pi_{GC}\alpha_{42}$	$\pi_{UA}\alpha_{54}$	$\pi_{UG}\alpha_{65}$	$\pi_{CG}\alpha_{75}$	$\pi_{AA}\alpha_{84}$	$\pi_{AG}\alpha_{92}$	$\pi_{AC}\alpha_{99}$	$\pi_{GA}\alpha_{105}$	$\pi_{GG}\alpha_{110}$	$\pi_{CA}\alpha_{114}$	$\pi_{CC}\alpha_{117}$	$\pi_{CU}\alpha_{119}$	$\pi_{UC}\alpha_{120}$

Figure 10. Definition of the rate matrix for the most general time reversible 16-state model of RNA evolution. π depicts the frequency parameter and α depicts the rate parameter. MM = mismatch, or 10 possible non-Watson-Crick basepairs. The four Watson-Crick basepairs and GU UG intermediates are bolded.

Table 7. Proposed RNA maximum likelihood models (modified from Savill *et al.*, 2001).

Model ^a	π parameters	r parameters ^b	model constraints ^c	free parameters ^d	comments
16 ¹	16: $\pi_1, \pi_2, \dots, \pi_{16}$	120: α_{ij}	2	134	- most GTR 16-state model.
16A ²	16: $\pi_1, \pi_2, \dots, \pi_{16}$	5: $\alpha_s, \alpha_d, \beta, \gamma, \epsilon$	2	19	- same parameters as model 7D, with $\epsilon =$ single substitutions \leftrightarrow mismatch (MM).
16B ³	16: $\pi_1, \pi_2, \dots, \pi_{16}$	1: μ	2	15	- $r_{ij} = \mu\pi_j$ when i and j differ by one substitution; $r_{ij} = 0$ otherwise.
16C ²	7: $\pi_1, \dots, \pi_6, \pi_m$	5: $\alpha_s, \alpha_d, \beta, \gamma, \epsilon$	2	10	- model 16A with all MM = π_m .
16O ⁴	4: $\pi_A, \pi_C, \pi_G, \pi_U$	8: $\alpha, \beta_1, \beta_2, S_{(AC)},$ $S_{(GU)}, S_{(CU)}, S_{(GA)}, S$	2	10	- modified K3ST81 (Kimura, 1981); 2 tv parameters; all MM = 0 ($S_{(MM)}$).
16D ²	4: $\pi_A, \pi_C, \pi_G, \pi_U$	4: $\alpha, \beta, \gamma, \phi$	2	6	- modified HKY (Hasegawa et al., 1985); rates for ts \neq tv; no bp reversal symmetry; G·U, U·G = ϕ (intermediate frequency).
16E ⁵	4: $\pi_A, \pi_C, \pi_G, \pi_U$	3: α, β, γ	2	5	- 16D with G·U, U·G = MM.
16F ⁵	4: $\pi_A, \pi_C, \pi_G, \pi_U$	3: α, β, γ	2	5	- 16E with G·U and U·G = Watson-Crick pairs.
16G ⁶	0	3: μ, β, γ	1	2	- model 16B with all frequencies equal.
16H ³	0	2: μ, γ	1	1	- simplest 16-state model.
7A ⁷	7: $\pi_1, \pi_2, \dots, \pi_7$	21: α_{ij}	2	26	- most GTR 7-state model.
7B ²	4: $\pi_1, \pi_2, \pi_3, \pi_7$	21: α_{ij}	2	23	- assumes bp reversal symmetry; i.e., $\pi_4 = \pi_1, \pi_5 = \pi_2, \pi_6 = \pi_3$.

Table 7. Continued.

Model ^a	π parameters	r parameters ^b	model constraints ^c	free parameters ^d	comments
7C ²	7: $\pi_1, \pi_2 \dots \pi_7$	10: α_{ij}	2	15	- double substitutions (ds) = 0; substitutions (s) \leftrightarrow mismatch (MM) \neq 0.
<u>7D</u> ⁸	7: $\pi_1, \pi_2 \dots \pi_7$	4: $\alpha_s, \alpha_d, \beta, \gamma$	2	9	- α_s = single substitutions (ss), α_d = ds, β = double transversions, γ = s \leftrightarrow MM.
7E ⁸	7: $\pi_1, \pi_2 \dots \pi_7$	2: α_s, γ	2	7	- model 7D with α_d = 0.
<u>7F</u> ²	4: $\pi_1, \pi_2, \pi_3, \pi_7$	4: $\alpha_s, \alpha_d, \beta, \gamma$	2	6	- model 7D with bp reversal symmetry.
6A ²	6: $\pi_1, \pi_2 \dots \pi_6$	15: α_{ij}	2	19	- most GTR 6-state model.
6B ²	6: $\pi_1, \pi_2 \dots \pi_6$	4: $\alpha_s, \alpha_d, \beta$	2	9	- model 7D without MM.
6C ⁹	3: π_1, π_2, π_3	4: $\alpha_s, \alpha_d, \beta$	2	6	- model 7F without MM.
6D ⁹	3: π_1, π_2, π_3	2: α_s, β	2	3	- model 6C with α_d = 0.
TN93 ¹⁰	4: $\pi_1, \pi_2 \dots \pi_4$	3: α_1, α_2	2	5	- model used on non-pairing regions

^a model names are followed from Savill *et al.* (2001) where applicable. Underlined models distinguish those which allow double substitutions across an RNA helix. Bold models distinguish those currently implemented in PHASE (Jow *et al.*, 2002; Hudelot *et al.*, 2003).

¹ Jow & Gowri-Shankar (2003), ² Savill *et al.* (2001), ³ Schöniger & von Haeseler (1994), ⁴ Otsuka *et al.* (1999), ⁵ Muse (1995),

⁶ Rhetsky (1995), ⁷ Higgs (2000), ⁸ Tillier & Collins (1998), ⁹ Tillier (1994), ¹⁰ Tamura & Nei (1993).

^b MM = mismatch (A•A, A^oG, A•C, C•A, C•C, C•U, G^oA, G•G, U•C, U•U); $S_{(AC)} = A•C, C•A$; $S_{(GU)} = G•U, U•G$; $S_{(CU)} = C•U, U•C$; $S_{(GA)} = G^oA, A^oG$; $S = A•A, U•U, G•G, C•C$; α_s = single transitions, α_d = double transitions, ts = transitions, tv = transversions, β = double transversions, γ = all substitutions to and from MM; ϵ = single substitutions to and from MM; ϕ = G•U, U•G pairs as intermediates; μ = scaling factor applied to $r_{ij} = \pi_j$ (Schöniger & von Haeseler, 1994) to satisfy constraint of substitution event per basepair in 1 unit time.

^c constraints posed for equilibrium frequencies (= to 1) and substitution events per base per unit time (= to 1).

^d no. free parameters = no. π parameters + no. r parameters - no. constraints.

substitutions to and from non-canonical pairs (Table 7). While having the most free parameters (number of rate and frequency parameters minus number of model constraints) of all simplifications of the 16-state models, Savill *et al.* (2001) demonstrated that model 16A outperforms simpler models that assume base-pair symmetry and a zero rate of double substitutions (models in Table 7). Model 16C of Savill *et al.* (2001) simplifies model 16A by assigning the same frequency parameter to all non-canonical pairs. Model 16D of Savill *et al.* (2001) is a modification of the HKY model of Hasegawa *et al.* (1985) that includes GU and UG pairs with a separate intermediate frequency. Other 16-state models and their characteristics are described in Table 7.

Seven-state models of RNA evolution assign all non-canonical base-pairs to one class, commonly called the mismatch class (Table 7). Tillier & Collins (1998) introduced the first seven-state models, model 7D and model 7E. Model 7D has four rate classes, single substitutions, double substitutions, double transversions, and changes to and from non-canonical pairs. Hence, it is similar to model 16A, except that it includes additional parameters for single substitutions to and from non-canonical pairs. Model 7E is a simplified version of model 7D in that all double substitutions are assigned a zero rate. Higgs (2000) produced the general time reversible seven-state model (7A), which has 21 individual rate classes and 26 free parameters. Despite also offering models 7B, 7C and 7F, all nested simplifications within model 7A (Table 7), Saville *et al.* (2001) demonstrated that model 7A outperforms other seven-state models that assume basepair symmetry or a zero rate of double substitutions.

Six-state models of RNA evolution do not consider any non-Watson-Crick base-pairs in their models, other than GU and UG intermediates (Table 7). Two of these models, 6C and 6D, were among the first proposed RNA likelihood models (Tillier, 1994). Model 6C assigns three rate parameters: single substitutions, double substitutions, and double transversions, while assuming basepair symmetry. Hence it is similar to model 7F except that it does not have a mismatch class. Model 6D simplifies model 6C by setting double transitions to a zero rate. Saville *et al.* (2001) introduced two more complex six-state models, 6A and 6B. Model 6A is the most general time reversible model for six-state models and has 19 free parameters (Table 7). Model 6B relaxes the assumptions of model 6A by assigning three rate parameters: single substitutions, double substitutions and double transversions.

Two software packages have incorporated some of the above-mentioned models into their programs. MrBayes ver 3.1 (and earlier versions) (Ronquist & Huelsenbeck, 2003) includes model 16B (Schöniger & von Haeseler, 1994) and allows for helices to be modeled independently as pairs along with other models for non-paired sites (i.e., loops, codons, amino acids). Importantly, model 16B should be considered a F81-like model for pairing sites, and when the covarion model in MrBayes is set to REV or HKY85, model 16B becomes different for each case (Jow, Gowri-Shankar & Guillard, unpublished PHASE ver. 2.0 manual). The program PHASE ver. 1.1 (Jow *et al.*, 2002) also provides a means to simultaneously model multiple partitions with different models of evolution. However, PHASE additionally contains a suite of RNA models that allow for the evaluation of the performance of different RNA models on a given dataset.

Likely as a result of the study of Saville *et al.* (2001), those models that allow for base-pair asymmetry and a non-zero rate of double substitutions, namely models 16A, 7A, 7D, 6A, and 6B, are all included in the PHASE program. Thus, PHASE allows the user to determine the best model of evolution for an RNA dataset rather than settle for only one RNA model (perhaps with slight modifications), as currently provided in MrBayes.

Phylogenetic studies that simultaneously incorporate RNA and DNA models for the analysis of rRNA and tRNA sequences are beginning to appear (e.g., Hudelot *et al.*, 2002; Jow *et al.*, 2003; Kjer, 2004; Gibson *et al.*, 2005; Gillespie *et al.*, 2005). However, none have provided a rigorous analysis of the variety of RNA models, as done by Saville *et al.* (2001), while simultaneously analyzing the non-pairing regions of the RNA-encoding sequences with a standard DNA model of substitution. In this study, I evaluate the existing suite of RNA maximum likelihood models provided in the PHASE ver. 1.1 program using a structural alignment (Gillespie *et al.*, 2004b) of the D2 and D3 expansion segments from 231 chrysomelid leaf beetles (Insecta: Coleoptera) (Fig. 11). These models are evaluated based on their performance over sampling generations of three and six million iterations using Markov Chain Monte Carlo simulation (Metropolis *et al.*, 1953; Hastings, 1970; Larget & Simon, 1999) under Bayesian analysis. The results are meaningful for practical phylogenetic studies including RNA sequences as markers.

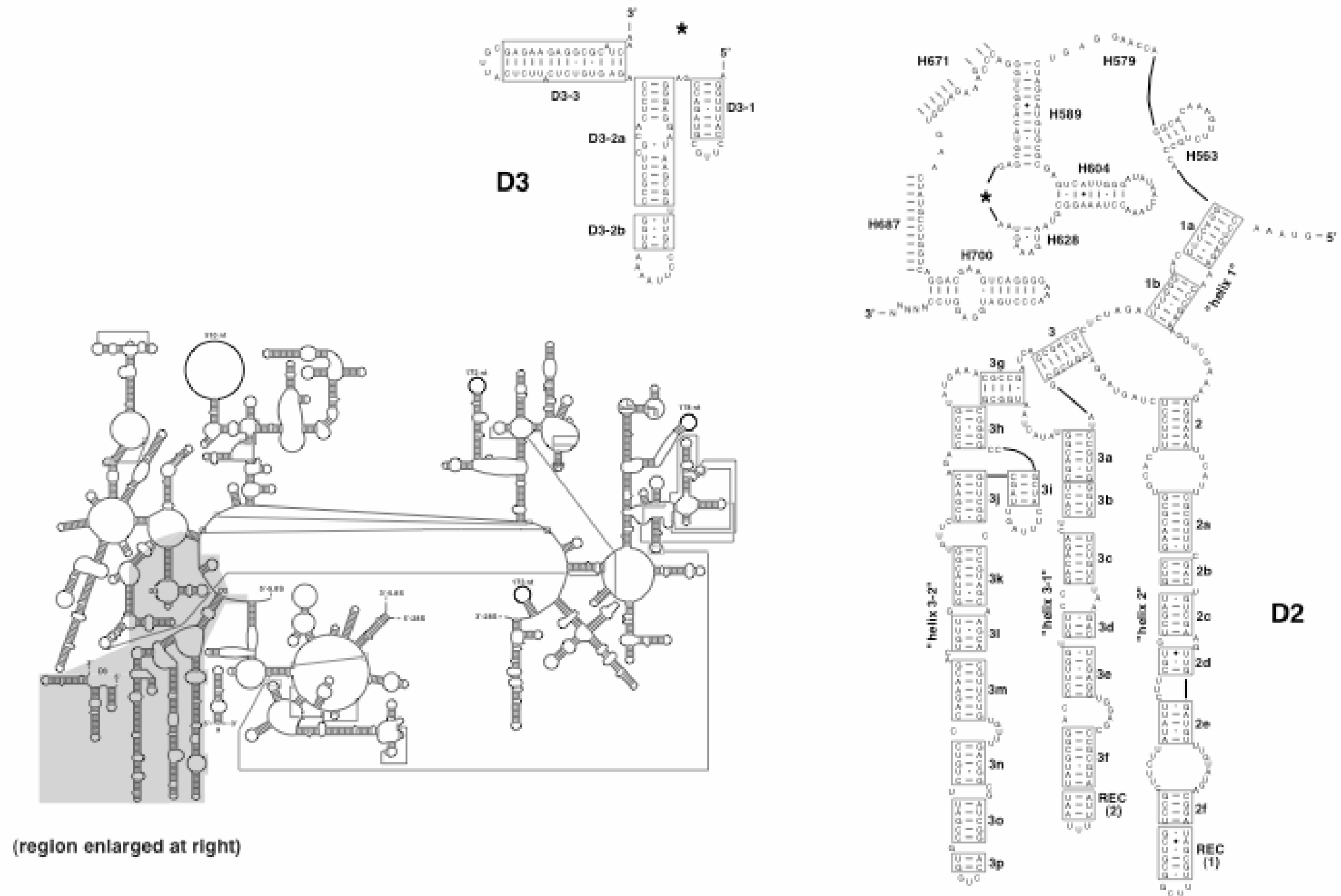


Figure 11. Predicted secondary structure model of the expansion segments D2 and D3 of the large subunit ribosomal RNA 28S of 231 sampled chrysomelid leaf beetles. The helices within squares were modeled under RNA likelihood models of evolution, while the unpaired sites were modeled using standard DNA maximum likelihood models of substitution. Unalignable regions, as described by Gillespie, and were not included in this analysis (See Gillespie *et al.* (2004b) for the location of these regions). Redrawn from Gillespie *et al.*, 2004b..

Results and discussion

Model performance

The log-likelihood values and AIC (Aikake Information Criterion, see Experimental procedures) calculations for the best and mean likelihood for the six- and seven-state models are provided in Table 8. Results from the 16-state models, for which analyses run for three million generations completed in just under one month, are not mentioned further given that the likelihoods and most model parameters did not reach stationarity. It is likely that the 16-state models are intractable for analyzing large numbers of sequences for long sampling procedures, unless many generations, likely over 10 million, are completed. For the six-state models, model 6A had the highest mean and best likelihoods after 3 million sampling generations, but higher AIC values due to a greater number of free parameters. This result was consistent after 6 million generations, except that the best likelihood for model 6B was better than the best likelihood for model 6A. For the seven-state models, the best likelihood for model 7D was also higher than the best likelihood for model 7A after 3 million generations; however, after six million generations model 7A had the higher likelihood. Still, as with the six-state models, the AIC values after both samplings are lower for the simpler model, model 7D, due to the high number of free parameters penalizing model 7A. Similarly, in combining the results of both the three- and six-million generation analyses for both models, the highest likelihoods were reported for the most general models, whereas the simplified models had lower AIC values due to the high number of free parameters (Table 8).

Table 8. Maximum likelihood values and AIC values for the best models and mean scores of 3000 and 6000 samples from the posterior probability distribution¹.

Model	3 million				6 million			
	Best		Mean		Best		Mean	
	-ln <i>L</i>	AIC ²	-ln <i>L</i>	AIC	-ln <i>L</i>	AIC ²	-ln <i>L</i>	AIC
6A	<u>7978.8428</u>	7997.8428	<u>8046.5120</u>	8065.5120	7984.8300	8003.8300	<u>8044.3500</u>	8063.3500
6B	7980.7551	<u>7989.7551</u>	8048.7190	<u>8057.7190</u>	<u>7982.5404</u>	<u>7991.5404</u>	8048.0690	<u>8057.0690</u>
7A	9548.4958	9574.4958	<u>9603.5970</u>	9629.5970	<u>9522.5937</u>	9548.5937	<u>9596.1280</u>	9622.1280
7D	<u>9540.4888</u>	<u>9549.4888</u>	9604.6300	<u>9613.6300</u>	9533.0801	<u>9542.0801</u>	9605.8210	<u>9614.8210</u>
16	-----	-----	-----	-----	-----	-----	-----	-----
16A	-----	-----	-----	-----	-----	-----	-----	-----
Model	Combined Mean							
	-ln <i>L</i>	AIC ²						
6A	<u>8046.0300</u>	8065.0300						
6B	8048.2720	<u>8057.2720</u>						
7A	<u>9598.4620</u>	9624.4620						
7D	9605.4480	<u>9614.4480</u>						
16	-----	-----						
16A	-----	-----						

¹ Underscored values depict models with better scores as compared to the other model in its class.
² Calculated as $AIC = -\ln L + \text{number of model free parameters}$. See Table 7 for model free parameters.

The mean likelihoods for all six- and seven-state likelihood models did not improve much when comparing the results from three million generations to those from six million generations (Table 8). This, however, does not imply that the best likelihoods were sampled in the posterior distribution, as multiple likely suboptimal likelihoods could be sampled in independent analyses. To address this issue, model mean parameter values and their estimated sampling size (ESS) were compared for all models over the two sampling generations (Tables 9-12). The mean likelihood for Model 6A has a low ESS after three million generations, but improves significantly after six million generations (Table 9). The likelihood plot of both analyses for model 6A suggests that even a highly conserved burn-in of 500,000 generations was not enough, and likely should be set to one million (Fig. 12). Setting the burn-in at 1,000,000 generations in Tracer resulted in a worse ESS (49.5) and further suggests that model 6A run for three million generations does not provide an efficient ESS for the mean likelihood. A sufficient sampling size is reached for the likelihood of model 6A over six million generations (Table 9, Fig. 12), however, four model parameters do not improve in their ESS (Table 3). Plots of these parameters (Fig. 13) indicate three of the four parameters (17, 21 and 30) got poorer in their ESS from three million to six million generations, suggesting that these parameters are not close to reaching convergence even after six million generations. Similarly, combining the results from the three- and six-million generation analyses only slightly increased the ESS for these parameters (Table 9).

The parameters of simpler six-state model, model 6B, reached stationarity after

Table 9. Mean and estimated samples sizes for model 6A statistics¹.

Statistic	3 million		6 million		Combined ²	
	Mean	ESS	Mean	ESS	Mean	ESS
lnLk	-8046.512	76.325	-8044.35	206.228	-8046.03	258.394
tree_prior	1579.225	7.591	1592.652	24.49	1576.218	31.356
TL	5.188	8.274	5.199	25.107	5.214	32.523
param0	0.444	198.197	0.446	358.237	0.443	536.661
param1	0.58	501.858	0.569	880.394	0.57	1337.966
param2	0.425	692.694	0.424	1136.249	0.425	1898.283
param3	0.169	594.851	0.171	1064.624	0.17	1720.96
param4	0.157	495.777	0.157	584.715	0.157	1048.04
param5	0.249	557.281	0.247	1017.943	0.248	1540.818
param6	0.36	140.101	0.353	312.155	0.356	453.368
param7	1.687	146.563	1.676	279.238	1.692	409.741
param8	1.804	114.135	1.766	283.125	1.772	400.21
param9	0.272	380.558	0.274	949.877	0.272	1336.765
param10	0.685	620.065	0.695	1100.893	0.691	1750.441
param11	0.208	110.057	0.21	229.702	0.209	347.72
param12	8.735E-2	467.217	8.726E-2	411.071	8.718E-2	895.342
param13	0.248	123.522	0.248	284.233	0.248	531.257
param14	0.182	181.81	0.184	257.402	0.184	455.396
param15	6.899E-2	205.706	6.945E-2	232.278	6.949E-2	452.572
param16	0.206	227.221	0.201	394.229	0.203	586.718
param17	1783.697	18.822	1636.488	8.24	1680.111	25.705
param18	29.498	472.842	25.414	598.716	26.591	1074.056
param19	64.483	111.57	54.199	326.096	57.153	443.422
param20	13.644	538.847	13.49	617.866	13.496	1197.302
param21	1192.432	14.551	1113.269	9.358	1134.23	22.548
param22	17.528	557.116	18.123	928.237	17.942	1539.73

Table 9 Continued.

Statistic	3 million		6 million		Combined²	
	Mean	ESS	Mean	ESS	Mean	ESS
param23	98.133	50.257	93.2	108.785	93.759	157.417
param24	10.647	331.585	10.687	1286.085	10.623	1650.851
param25	26.541	310.88	24.677	766.469	25.135	1086.644
param26	25.868	422.194	23.886	562.666	24.392	1016.201
param27	1.926	543.887	1.764	983.339	1.807	1570.553
param28	1145.754	6.939	1064.705	9.204	1081.247	16.601
param29	121.312	94.658	113.398	185.289	115.396	282.581
param30	903.271	9.453	867.267	9.17	871.253	18.081

¹ Bold values depict likelihoods and parameters with a poor estimated sample size.

² Results from both analyses combined in Tracer.

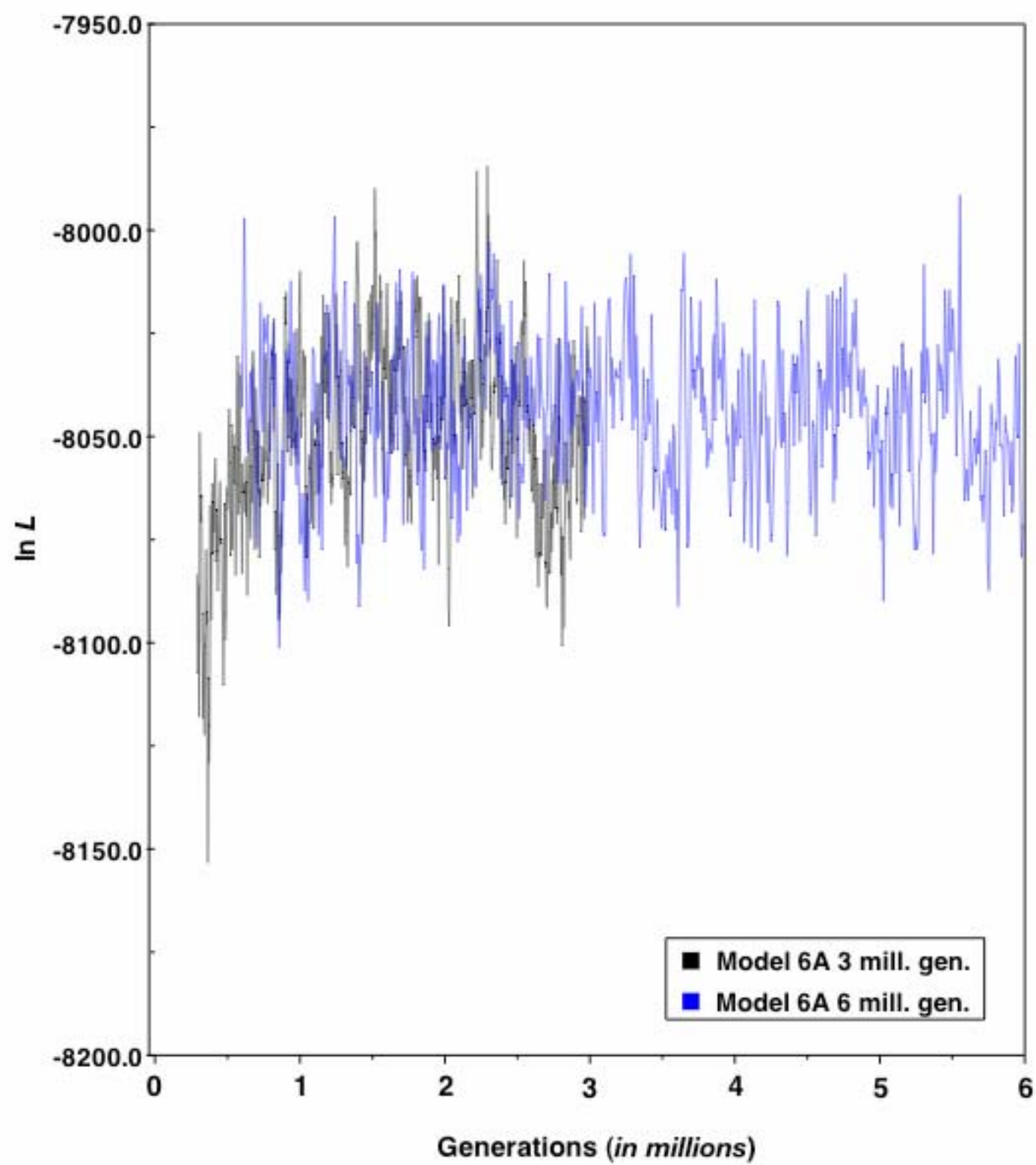


Figure 12. Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 6A.

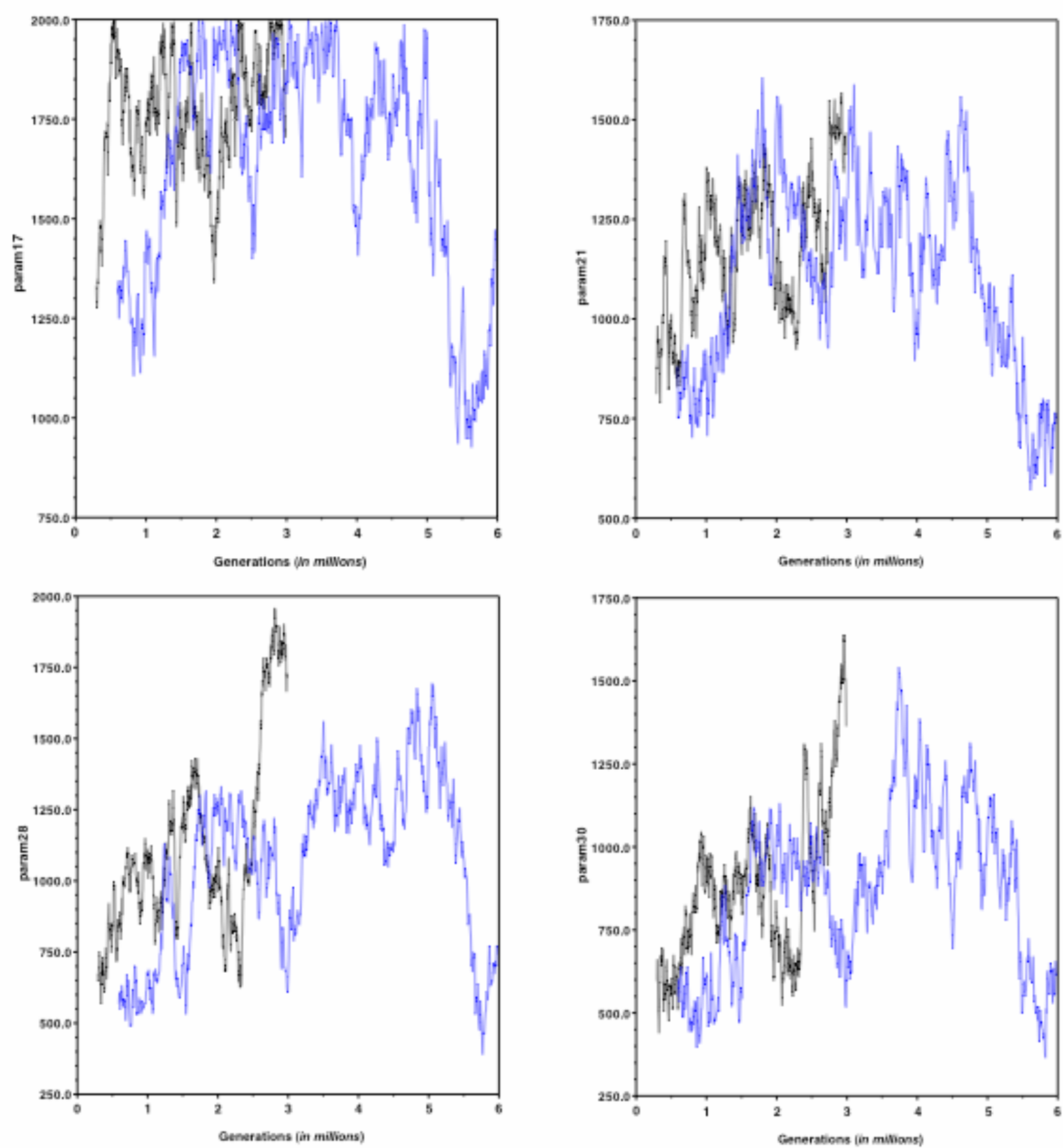


Figure 13. Plots of the four poor parameters over three and six generations for the analyses performed under model 6A.

Table 10. Mean and estimated samples sizes for model 6B statistics¹.

Statistic	3 million		6 million		Combined ²	
	Mean	ESS	Mean	ESS	Mean	ESS
lnLk	-8048.719	88.639	-8048.069	143.178	-8048.272	231.818
tree_prior	1586.219	14.253	1566.315	11.005	1572.535	25.258
TL	5.193	18.245	5.302	18.695	5.268	36.94
param0	0.444	266.76	0.436	590.453	0.439	857.213
param1	0.572	709.79	0.552	1065.234	0.558	1775.025
param2	0.425	856.648	0.427	1543.96	0.427	2400.608
param3	0.166	537.462	0.17	1421.009	0.169	1958.471
param4	0.155	715.502	0.155	1551.268	0.155	2266.77
param5	0.253	557.282	0.248	1319.667	0.25	1876.949
param6	0.345	332.852	0.349	742.553	0.348	1075.405
param7	1.629	268.496	1.631	681.173	1.63	949.668
param8	1.534	204.462	1.551	128.282	1.546	332.744
param9	0.293	773.358	0.304	1909.784	0.301	2683.142
param10	0.79	778.521	0.809	2248.192	0.803	3026.713
param11	0.23	1046.39	0.231	2002.459	0.231	3048.849
param12	0.116	868.302	0.114	1049.397	0.114	1917.699
param13	0.195	1077.534	0.199	2282.955	0.198	3360.489
param14	0.203	1450.668	0.205	1971.489	0.205	3422.157
param15	7.599E-2	511.753	7.294E-2	499.444	7.39E-2	1011.197
param16	0.179	1217.328	0.178	1964.986	0.179	3182.314
param17	8.302	500.028	8.788	689.573	8.636	1189.6
param18	0.151	818.714	0.156	1884.154	0.154	2702.868

¹ Bold values depict values with a poor estimated sample size.

² Results from both analyses combined in Tracer.

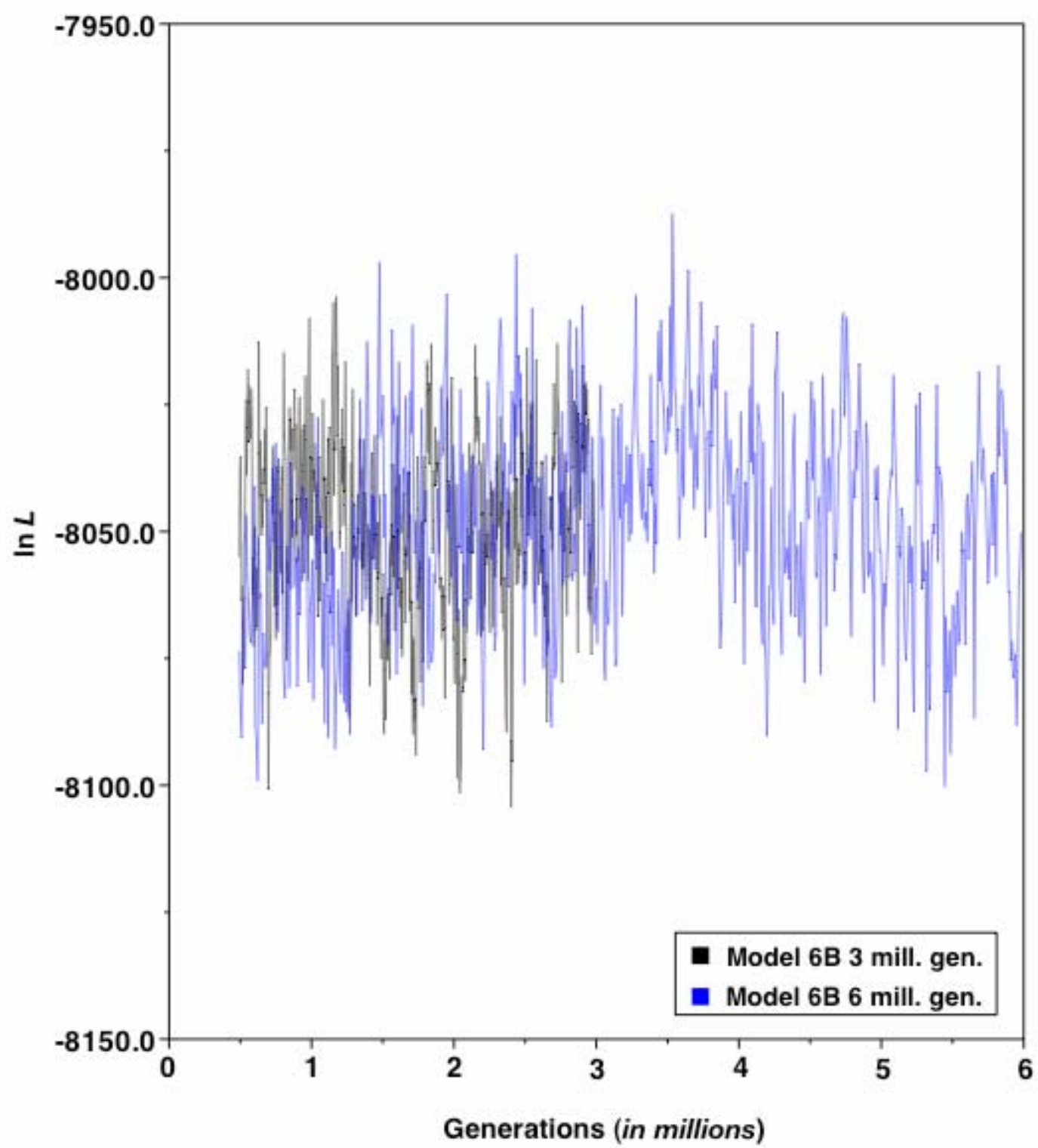


Figure 14. Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 6B.

three million generations with sufficient ESS for each (Table 10). The ESS for the mean likelihood was poor after three million generations but increased greatly after six million generations (Table 10), with plots of both mean likelihoods suggesting convergence on all model 6B statistics (Fig. 14). A burn-in of 500,000 generations was apparently enough for the simpler six-state model.

The mean likelihood for Model 7A has a sufficient ESS after three million generations, and improves after six million generations (Table 11). The likelihood plot of both analyses for model 7A suggests a burn-in of 500,000 generations was enough and that stationarity was reached for the mean likelihood (Fig. 15). Like the most general six-state model, certain parameters of model 7A did not reach convergence even after six million generations (Fig. 16), with poor ESS values suggesting many more generations would be needed for these parameters to reach stationarity (Table 11). However, unlike model 6A, wherein three of the four poor parameters did not increase in ESS from three million to six million generations, the ten poor parameters of model 7A all increased in ESS from three million to six million generations, and further increased when means from both analyses were combined (Table 11). Given that all of the poor parameters are in rate classes with low frequencies (those modeling non-Watson-Crick base-pairs), it is likely that non-canonical base-pairs are more difficult to model due to their infrequency in the dataset. Still, an increase in ESS over longer generation times suggests that, at some point, even these additional parameters will reach stationarity.

The ESS for the mean likelihood for model 7D improved over six million generations (Fig. 17) but did not reach a sufficient level until both analyses were

Table 11. Mean and estimated samples sizes for model 7A statistics¹.

Statistic	3 million		6 million		Combined ²	
	Mean	ESS	Mean	ESS	Mean	ESS
lnLk	-9603.597	107.427	-9596.128	150.725	-9598.462	258.151
tree_prior	1525.508	5.381	1591.061	26.266	1570.576	31.647
TL	5.498	15.652	5.155	30.138	5.262	45.79
param0	0.438	216.564	0.438	275.889	0.438	492.453
param1	0.571	449.448	0.564	664.617	0.566	1114.065
param2	0.426	495.261	0.426	876.488	0.426	1371.749
param3	0.173	406.269	0.171	756.609	0.172	1162.877
param4	0.156	292.819	0.159	421.862	0.158	714.681
param5	0.246	353.926	0.244	513.522	0.244	867.448
param6	0.357	128.181	0.374	214.591	0.369	342.772
param7	1.598	118.968	1.746	180.322	1.7	299.29
param8	1.984	110.996	2.148	167.15	2.096	278.146
param9	0.217	522.661	0.214	1006.379	0.215	1529.04
param10	0.821	516.157	0.816	1080.942	0.818	1597.098
param11	0.199	94.036	0.198	164.127	0.198	258.162
param12	8.384E-2	130.229	8.216E-2	254.553	8.268E-2	384.782
param13	0.233	112.435	0.231	221.931	0.232	334.365
param14	0.182	105.268	0.181	266.411	0.181	371.679
param15	5.925E-2	98.61	5.905E-2	200.581	5.911E-2	299.192
param16	0.171	85.471	0.177	248.154	0.175	333.625
param17	7.134E-2	54.849	7.192E-2	262.836	7.174E-2	317.685
param18	1490.244	3.883	1609.128	6.035	1571.977	9.918
param19	11.72	295.226	12.514	671.853	12.266	967.079
param20	20.491	140.926	21.677	546.677	21.306	687.602
param21	6.102	88.659	5.776	383.55	5.878	472.209
param22	532.042	15.48	549.611	17.153	544.121	32.633

Table 11 Continued.

Statistic	3 million		6 million		Combined ²	
	Mean	ESS	Mean	ESS	Mean	ESS
param23	972.373	3.764	1050.047	6.404	1025.774	10.168
param24	7.958	637.097	9.02	581.807	8.688	1218.904
param25	53.951	86.02	49.64	246.173	50.987	332.193
param26	9.224	92.43	10.085	203.634	9.816	296.063
param27	384.25	8.603	354.994	27.418	364.137	36.021
param28	11.839	291.213	12.384	516.709	12.214	807.922
param29	13.338	162.684	13.551	414.328	13.485	577.011
param30	1.287	237.425	1.353	707.016	1.332	944.441
param31	425.198	14.222	450.753	30.486	442.767	44.708
param32	1068.735	3.662	1161.558	6.769	1132.551	10.431
param33	109.355	13.851	113.95	115.739	112.514	129.59
param34	646.973	2.953	681.531	16.372	670.731	19.325
param35	1040.78	2.944	1076.255	9.556	1065.169	12.5
param36	1306.581	3.234	1517.419	5.853	1451.532	9.086
param37	361.142	16.387	360.947	23.185	361.008	39.571

¹ Bold values depict values with a poor estimated sample size.

² Results from both analyses combined in Tracer.

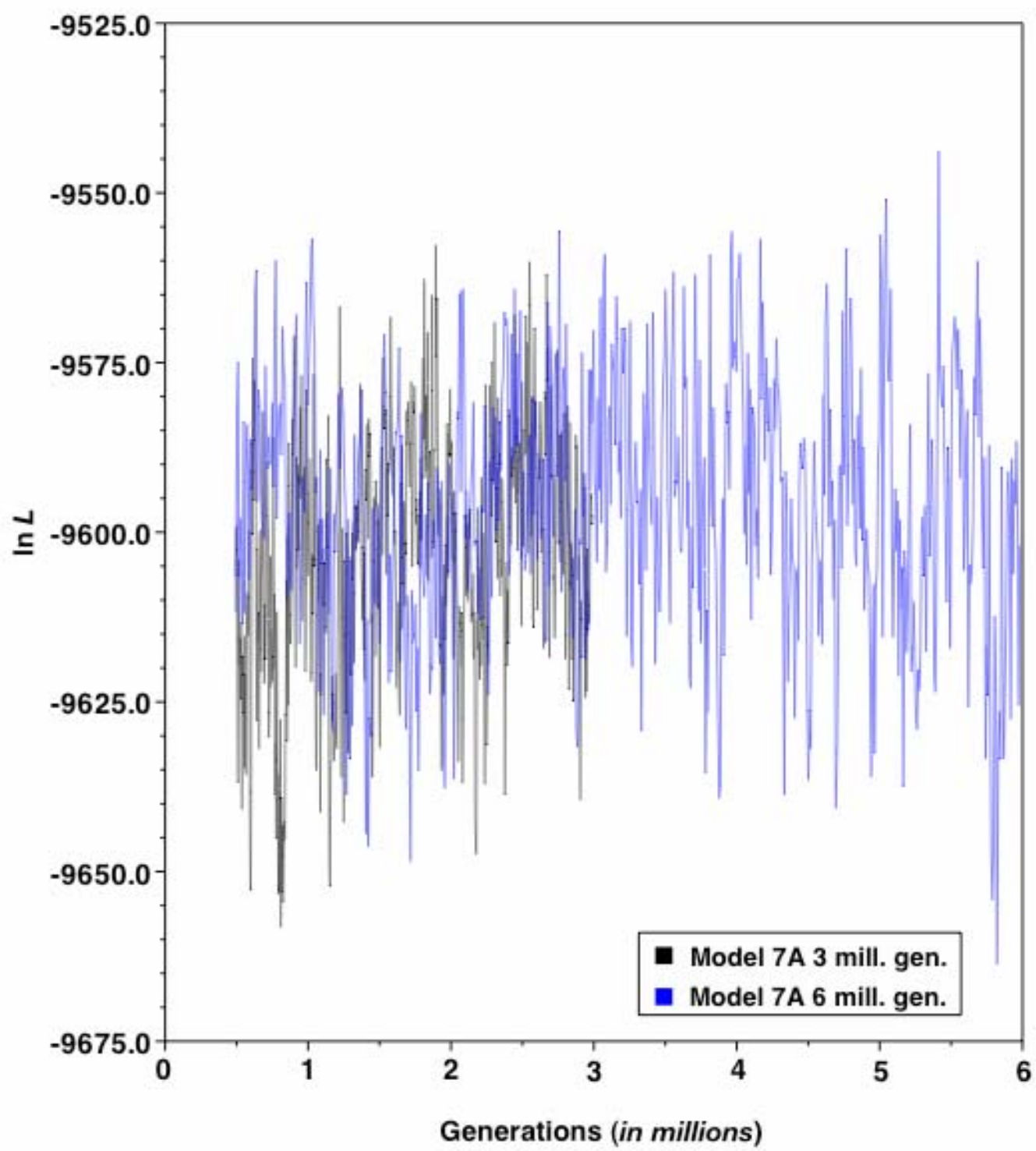


Figure 15. Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 7A.

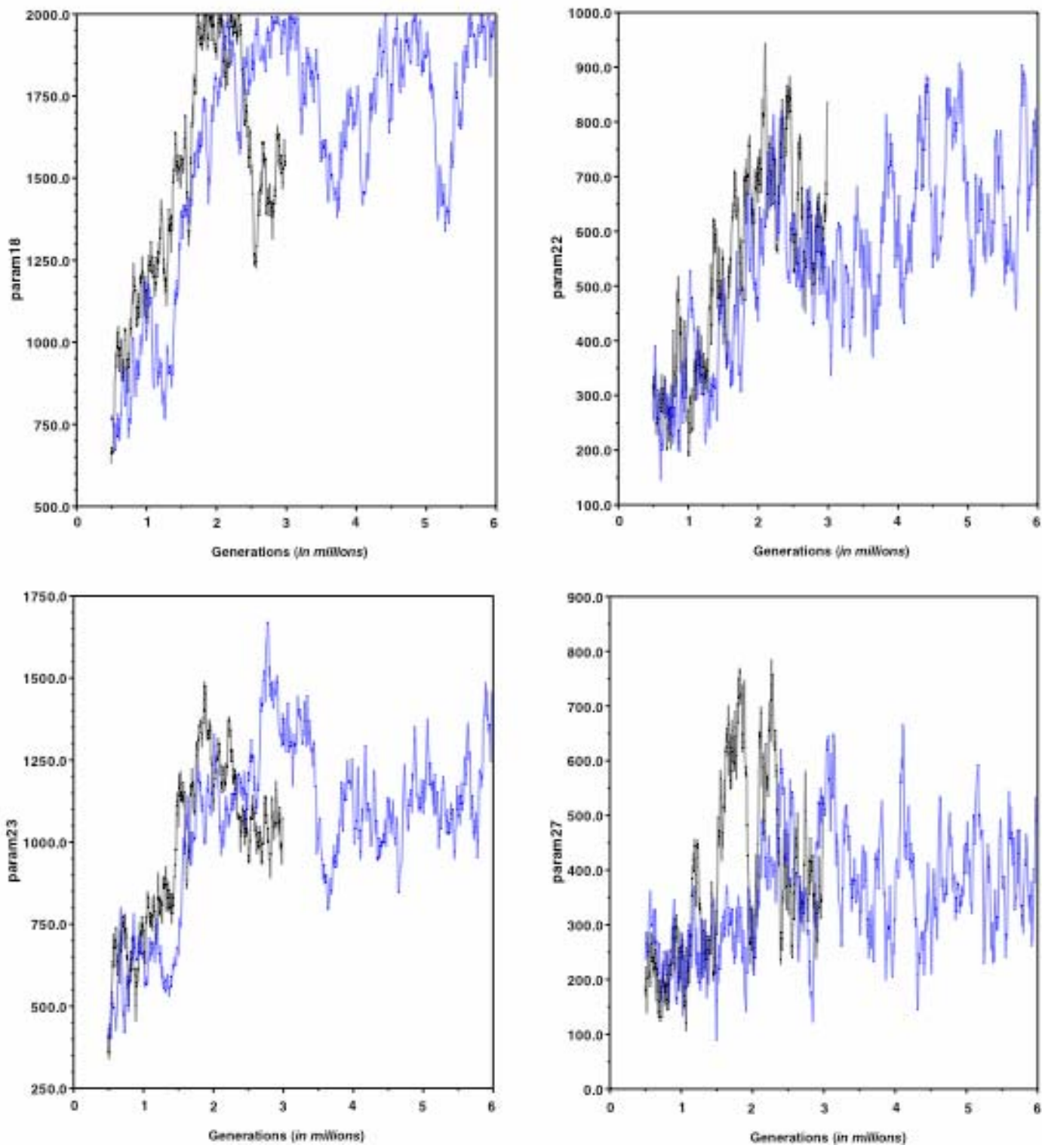


Figure 16. Plots of the ten poor parameters over three and six generations for the analyses performed under model 7A.

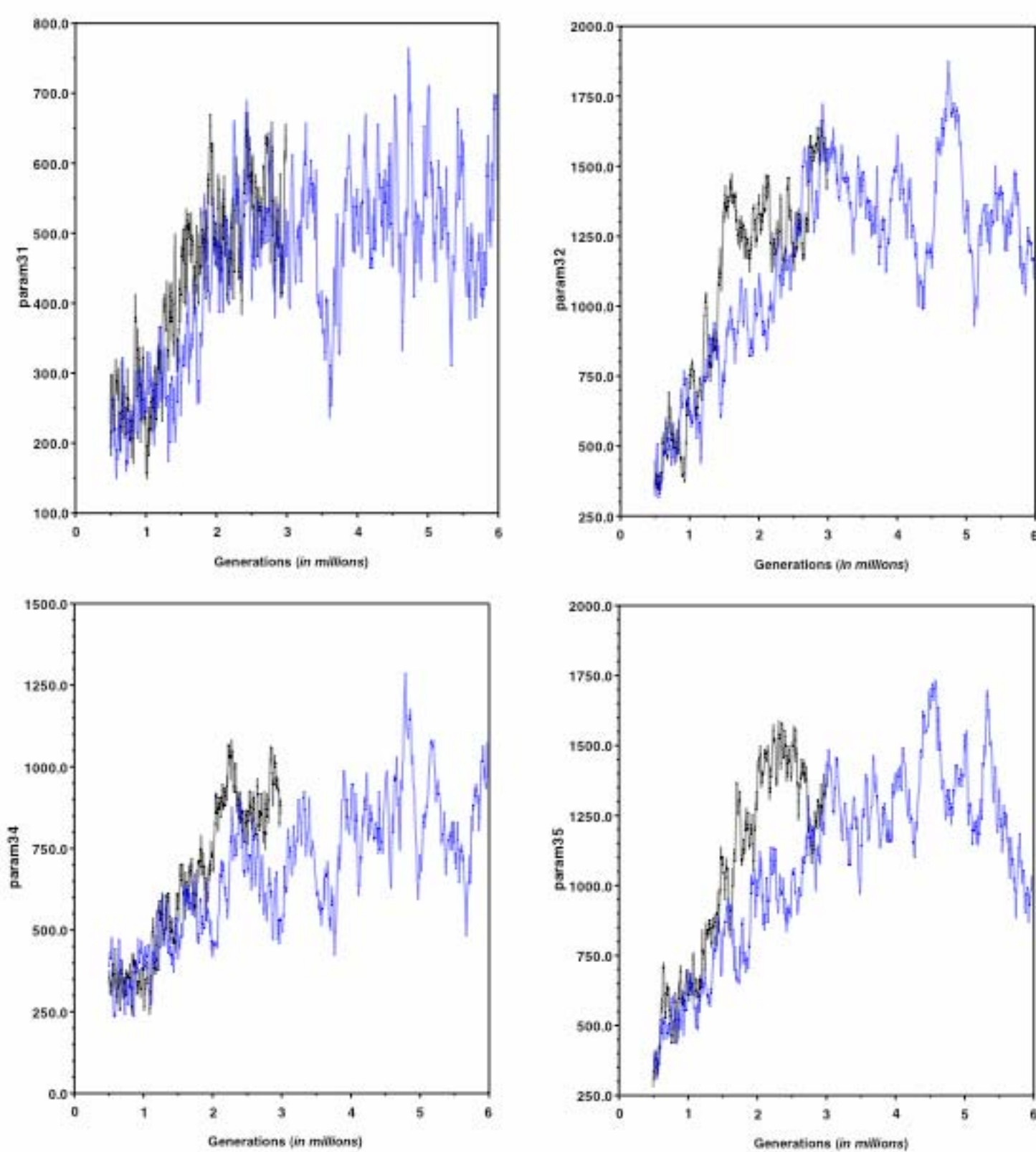


Figure 16 Continued.

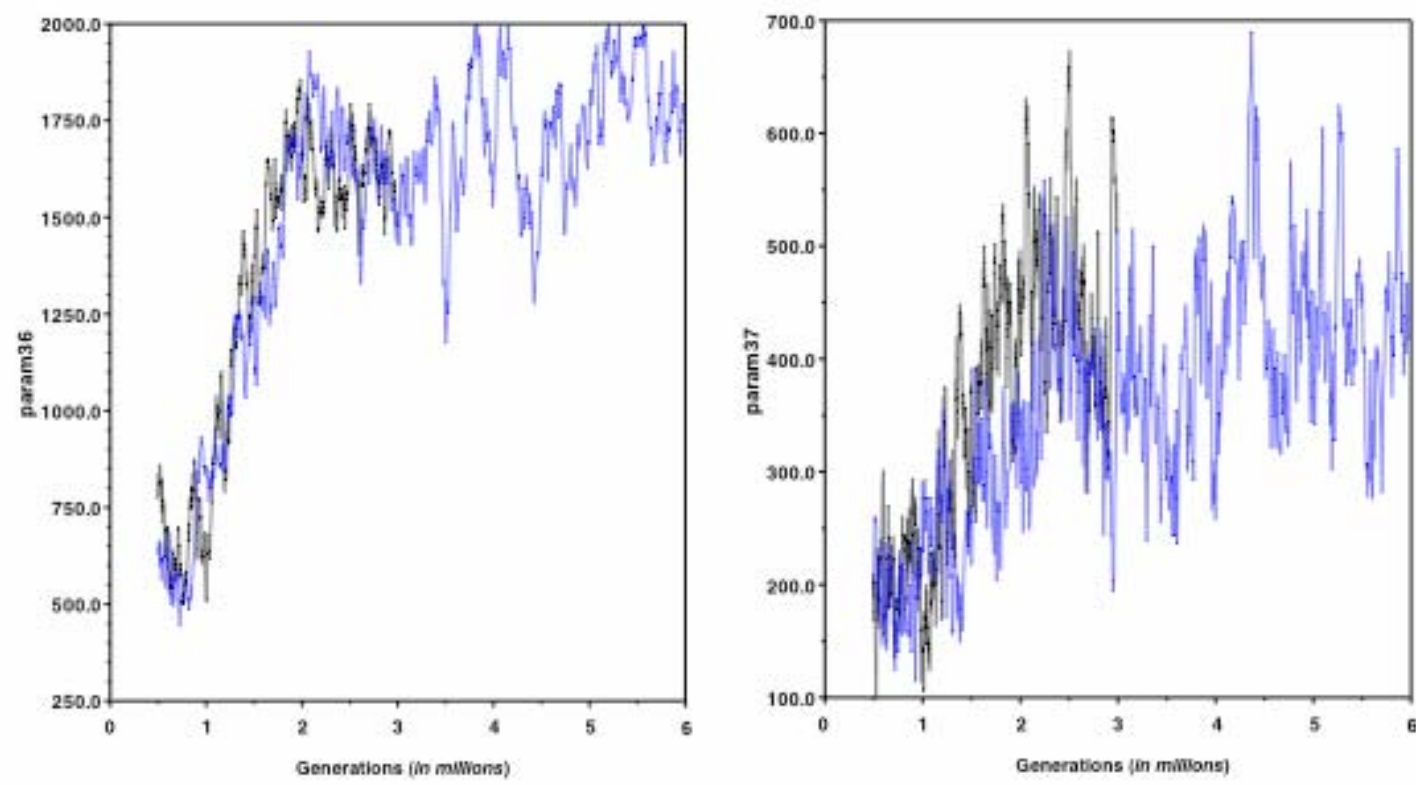


Figure 16 Continued.

Table 12. Mean and estimated samples sizes for model 7D statistics¹.

Statistic	3 million		6 million		Combined ²	
	Mean	ESS	Mean	ESS	Mean	ESS
lnLk	-9604.63	51.883	-9605.821	>99.888	-9605.448	151.772
tree_prior	1574.314	4.999	1584.752	>11.806	1581.49	>16.805
TL	5.04	15.898	5.144	>38.32	5.111	>54.219
param0	0.453	367.179	0.446	633.487	0.448	1000.665
param1	0.59	722.391	0.575	1356.74	0.58	2079.131
param2	0.43	654.462	0.427	1226.468	0.428	1880.93
param3	0.173	639.382	0.17	1021.268	0.171	1660.65
param4	0.154	489.572	0.156	843.932	0.156	1333.504
param5	0.243	659.646	0.246	1021.22	0.245	1680.866
param6	0.356	219.26	0.363	543.937	0.361	763.197
param7	1.574	260.214	1.645	467.156	1.623	727.37
param8	1.959	339.637	1.931	291.644	1.94	631.281
param9	0.225	1055.22	0.223	1238.728	0.224	2293.948
param10	0.86	1037.559	0.865	1529.714	0.863	2567.273
param11	0.221	825.985	0.22	1451.022	0.221	2277.007
param12	0.102	446.042	0.101	1438.775	0.102	1884.817
param13	0.183	786.456	0.184	2052.433	0.184	2838.889
param14	0.2	792.361	0.2	1613.223	0.2	2405.583
param15	6.982E-2	661.673	7.037E-2	1724.353	7.02E-2	2386.026
param16	0.159	976.362	0.159	1550.206	0.159	2526.568
param17	6.491E-2	321.98	6.493E-2	686.441	6.492E-2	1008.421
param18	9.393	259.334	9.458	611.297	9.438	870.631
param19	5.486E-2	1091.751	5.29E-2	2582.322	5.351E-2	3674.074
param20	6.005	210.855	6.104	494.611	6.073	705.466

¹ Bold values depict values with a poor estimated sample size.

² Results from both analyses combined in Tracer.

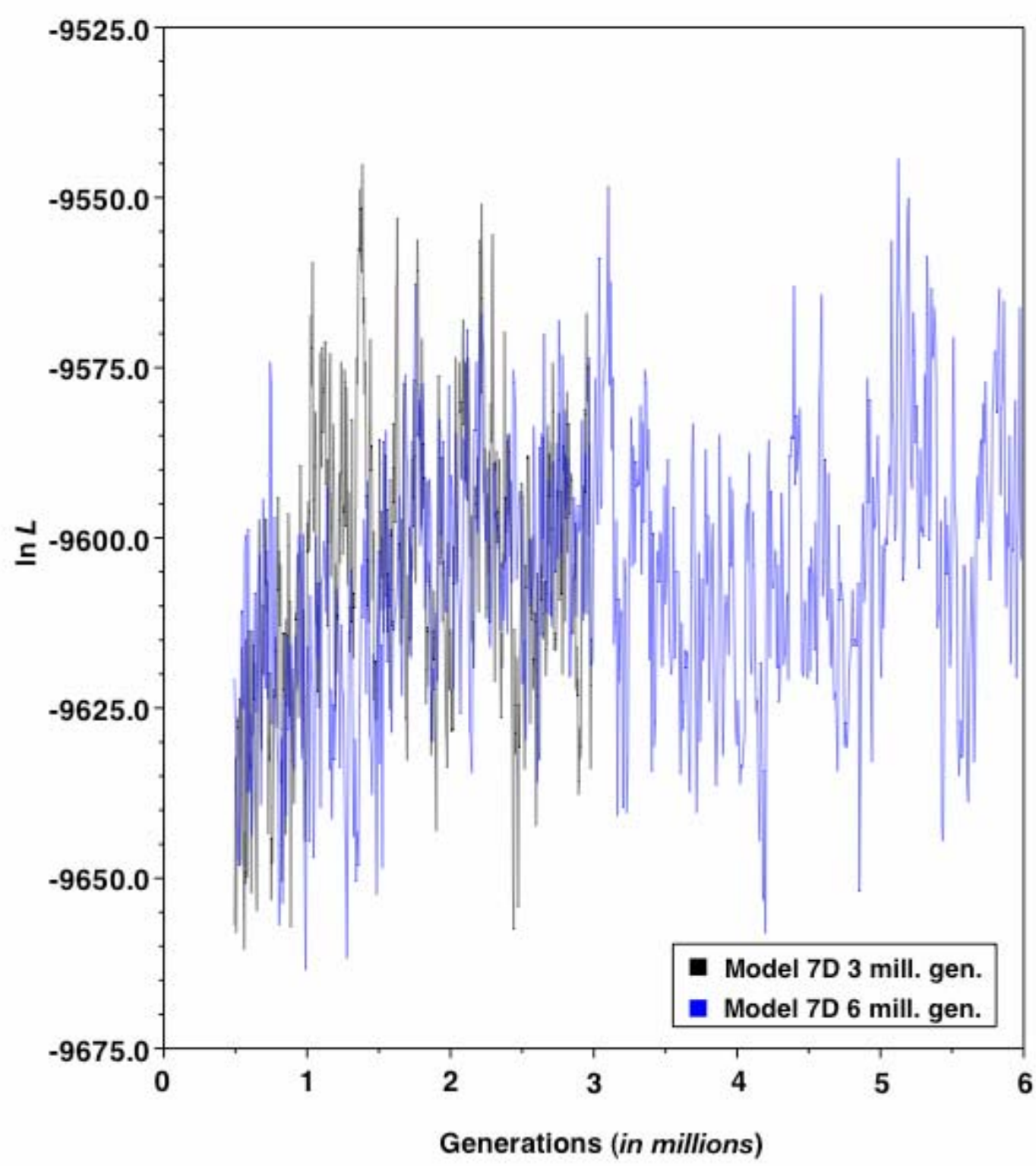


Figure 17. Plot of the log likelihood ($\ln L$) over the sampled generations for the three and six million generation analyses performed under model 7D.

combined (Table 12). Interestingly, all other model parameters reached convergence after three million generations, suggesting that model 7D was run for a sufficient number of generations. This also suggests that, when lumped into one rate class (γ), rare substitutions to and from non-canonical basepairs reach stationarity at a more rapid rate in parameter space.

RNA maximum likelihood model assumptions

Savill *et al.* (2001) demonstrated that RNA maximum likelihood models that do not assume base-pair reversal symmetry or zero rates for double substitutions perform better than those models that do. The rate matrices for the best likelihood scores for each model in this study illustrate the different effects of model parameters on the dataset (Table 13). It is clear that relaxing base-pair asymmetry in these models would not appropriately accommodate the pattern of nucleotide substitution in these data (using other models listed in Table 7), as no substitution class is identical in its basepair reversal symmetry. Thus, my results are similar to those of Savill *et al.* (2001) and suggest the continued use of models that do not assume basepair reversal symmetry.

Models that assume double substitutions occur across RNA helices make certain assumptions about the dynamics of compensatory base change in these molecules. Within populations of organisms, GU UG intermediates in RNA molecules can either occur in high or low frequencies as slightly deleterious mutations (or arguably stable GU UG basepairs that are favored over Watson-Crick base-pairs). If these intermediates occur at high frequency within the population, then a single transition at one position

Table 13. Best rate matrices of the models evaluated in this study.¹.

Model 6A							Model 6B						
	AU	GU	GC	UA	UG	CG		AU	GU	GC	UA	UG	CG
AU	-----	0.7770	0.0012	0.0269	0.0245	0.0169	AU	-----	0.6545	0.1047	0.0191	0.0056	0.0173
GU	2.1297	-----	1.8444	0.0035	0.0758	0.0181	GU	1.3572	-----	1.1866	0.0191	0.0056	0.0173
GC	0.0011	0.6165	-----	0.0388	0.0039	0.0000	GC	0.1198	0.6545	-----	0.0191	0.0056	0.0173
UA	0.0278	0.0013	0.0438	-----	0.3639	0.0967	UA	0.0198	0.0096	0.0173	-----	0.3851	0.1045
UG	0.0607	0.0685	0.0104	0.8712	-----	0.8334	UG	0.0198	0.0096	0.0173	1.3059	-----	1.1842
CG	0.0179	0.0070	0.0000	0.0994	0.3580	-----	CG	0.0198	0.0096	0.0173	0.1152	0.3851	-----

Model 7A								Model 7D							
	AU	GU	GC	UA	UG	CG	MM		AU	GU	GC	UA	UG	CG	MM
AU	-----	0.5857	0.0010	0.0070	0.0072	0.0010	0.1593	AU	-----	0.4369	0.0977	0.0064	0.0020	0.0040	0.1837
GU	1.5196	-----	1.0747	0.0017	0.0038	0.0044	0.0905	GU	0.9957	-----	0.7893	0.0064	0.0020	0.0040	0.1837
GC	0.0009	0.3754	-----	0.0081	0.0010	0.0004	0.1304	GC	0.1233	0.4369	-----	0.0064	0.0020	0.0040	0.1837
UA	0.0099	0.0009	0.0126	-----	0.3691	0.0740	0.2621	UA	0.0070	0.0031	0.0056	-----	0.2899	0.0711	0.1837
UG	0.0291	0.0060	0.0045	1.0644	-----	0.9945	0.3927	UG	0.0070	0.0031	0.0056	0.9124	-----	0.5745	0.1837
CG	0.0012	0.0019	0.0004	0.0594	0.2769	-----	0.1071	CG	0.0070	0.0031	0.0056	0.1129	0.2899	-----	0.1837
MM	0.5659	0.1239	0.5110	0.6608	0.3434	0.3364	-----	MM	0.6719	0.2948	0.5326	0.6156	0.1956	0.3876	-----

¹Calculated in PHASE ver. 1.1. Best rate matrices are rescaled so that the average substitution rate is 1.0.

would create a more stable Watson-Crick base-pair that would be energetically advantageous over the intermediate. If this new allele is the opposite of the dominant allele, it could be selectively swept through the population (Savill *et al.* (2001). Alternatively, if GU UG intermediates are always present in some low frequency within populations, then single transitions will always introduce alternatives to the dominant alleles in the population just by random drift in allelic frequencies due to natural processes (e.g., Kimura, 1985; Stephan, 1996; Higgs, 1998). Thus, from this viewpoint it is easy to see how double substitutions can become fixed within species (Savill *et al.* (2001).

In order to evaluate the occurrence of double substitutions in this dataset, I ranked the rate matrices from Table 13 according to rate class frequency (Tables 14-15). In the case of the six-state models, double substitutions account for 0.6442 and 0.7103 of the overall substitution frequency in models 6A and 6B, respectively. Similarly, double substitutions comprise 0.2364 and 0.4893 of the overall substitution frequency for seven-state models 7A and 7D, respectively. Given this, it seems logical to parameterize double substitutions; however, current simplifications of double substitution classes may be misleading. For example, double transitions are often assigned to one rate class, α_d . But in ranking the frequency of all substitution classes, it appears that double transitions with a pyrimidine on the 5'-side of the base-pair (CG→UA, UA→CG) occur in much greater frequency than double transitions with a purine on the 5'-side of the base-pair (AU→GC, GC→AU) (Tables 14-15). The cause of this asymmetry in double transitions is unknown, and to my knowledge, it has previously not been detected in studies on

Table 14. Rank of substitution types and frequencies estimated for six-state RNA models for basepairs within the helices of the 28S rRNA expansion segments and related core elements¹.

Model 6A			Model 6B		
Substitution	Type ²	Mutability	Substitution	Type ²	Mutability
GU→AU	α_s	2.1297	GU→AU	α_s	1.3572
GU→GC	α_s	1.8444	UG→UA	α_s	1.3059
UG→UA	α_s	0.8712	GU→GC	α_s	1.1866
UG→CG	α_s	0.8334	UG→CG	α_s	1.1842
AU→GU	α_s	0.7770	AU→GU	α_s	0.6545
GC→GU	α_s	0.6165	GC→GU	α_s	0.6545
UA→UG	α_s	0.3639	UA→UG	α_s	0.3851
CG→UG	α_s	0.3580	CG→UG	α_s	0.3851
CG→UA	α_d	0.0994	GC→AU	α_d	0.1198
UA→CG	α_d	0.0967	CG→UA	α_d	0.1152
GU→UG	β	0.0758	AU→GC	α_d	0.1047
UG→GU	β	0.0685	UA→CG	α_d	0.1045
UG→AU	β	0.0607	UA→AU	β	0.0198
UA→GC	β	0.0438	UG→AU	β	0.0198
GC→UA	β	0.0388	CG→AU	β	0.0198
UA→AU	β	0.0278	AU→UA	β	0.0191
AU→UA	β	0.0269	GU→UA	β	0.0191
AU→UG	β	0.0245	GC→UA	β	0.0191
GU→CG	β	0.0181	AU→CG	β	0.0173
CG→AU	β	0.0179	GU→CG	β	0.0173
AU→CG	β	0.0169	GC→CG	β	0.0173
UG→GC	β	0.0104	UA→GC	β	0.0173
CG→GU	β	0.0070	UG→GC	β	0.0173
GC→UG	β	0.0039	CG→GC	β	0.0173
GU→UA	β	0.0035	UA→GU	β	0.0096
UA→GU	β	0.0013	UG→GU	β	0.0096
AU→GC	α_d	0.0012	CG→GU	β	0.0096
GC→AU	α_d	0.0011	AU→UG	β	0.0056
GC→CG	β	0.0000	GU→UG	β	0.0056
CG→GC	β	0.0000	GC→UG	β	0.0056

¹ α_s = single transitions, α_d = double transitions, α_t = transitions, α_{tv} = transversions, β = double transversions, γ = all substitutions to and from MM

² Shaded values are discussed in the text.

nucleotide rate evolution in RNA datasets. Perhaps this is a property specific to rRNA expansion segment evolution, or even to this group of beetle taxa. Nevertheless,

Table 15. Rank of substitution types and frequencies estimated for seven-state RNA models for basepairs within the helices of the 28S rRNA expansion segments and related core elements ^{1,2}.

Model 7A			Model 7D		
Substitution	Type ²	Mutability	Substitution	Type ²	Mutability
GU→AU	α_s	1.5196	GU→AU	α_s	0.9957
GU→GC	α_s	1.0747	UG→UA	α_s	0.9124
UG→UA	α_s	1.0644	GU→GC	α_s	0.7893
UG→CG	α_s	0.9945	MM→AU	γ	0.6719
MM→UA	γ	0.6608	MM→UA	γ	0.6156
AU→GU	α_s	0.5857	UG→CG	α_s	0.5745
MM→AU	γ	0.5659	MM→GC	γ	0.5326
MM→GC	γ	0.5110	AU→GU	α_s	0.4369
UG→MM	γ	0.3927	GC→GU	α_s	0.4369
GC→GU	α_s	0.3754	MM→CG	γ	0.3876
UA→UG	α_s	0.3691	MM→GU	γ	0.2948
MM→UG	γ	0.3434	UA→UG	α_s	0.2899
MM→CG	γ	0.3364	CG→UG	α_s	0.2899
CG→UG	α_s	0.2769	MM→UG	γ	0.1956
UA→MM	γ	0.2621	AU→MM	γ	0.1837
AU→MM	γ	0.1593	GU→MM	γ	0.1837
GC→MM	γ	0.1304	GC→MM	γ	0.1837
MM→GU	γ	0.1239	UA→MM	γ	0.1837
CG→MM	γ	0.1071	UG→MM	γ	0.1837
GU→MM	γ	0.0905	CG→MM	γ	0.1837
UA→CG	α_d	0.0740	GC→AU	α_d	0.1233
CG→UA	α_d	0.0594	CG→UA	α_d	0.1129
UG→AU	β	0.0291	AU→GC	α_d	0.0977
UA→GC	β	0.0126	UA→CG	α_d	0.0711
UA→AU	β	0.0099	UA→AU	β	0.0070
GC→UA	β	0.0081	UG→AU	β	0.0070
AU→UG	β	0.0072	CG→AU	β	0.0070
AU→UA	β	0.0070	AU→UA	β	0.0064
UG→GU	β	0.0060	GU→UA	β	0.0064
UG→GC	β	0.0045	GC→UA	β	0.0064
GU→CG	β	0.0044	UA→GC	β	0.0056
GU→UG	β	0.0038	UG→GC	β	0.0056
CG→GU	β	0.0019	CG→GC	β	0.0056
GU→UA	β	0.0017	AU→CG	β	0.0040
CG→AU	β	0.0012	GU→CG	β	0.0040
AU→GC	α_d	0.0010	GC→CG	β	0.0040
GC→AU	α_d	0.0009	UA→GU	β	0.0031
GC→UG	β	0.0010	UG→GU	β	0.0031
AU→CG	β	0.0010	CG→GU	β	0.0031

Table 15 Continued.

Model 7A			Model 7D		
Substitution	Type ²	Mutability	Substitution	Type ²	Mutability
UA→GU	β	0.0009	AU→UG	β	0.0020
GC→CG	β	0.0004	GU→UG	β	0.0020
CG→GC	β	0.0004	GC→UG	β	0.0020

¹ α_s = single transitions, α_d = double transitions, α_t = transitions, α_{tv} = transversions, β = double transversions, γ = all substitutions to and from MM

² Shaded values are discussed in the text.

assymetries in the rates of double transitions suggest that models that do not specifically parameterize *each type* of double transition (in this study, models 6B and 7D) will grossly overestimate the rate at which AU→GC and GC→AU substitutions occur. The effects this will have on reconstructed phylogenies remains to be tested.

Model selection

Likelihood ratio tests (LRT) performed on the two models within each RNA model class provide some insight as to which performed better (Table 16). The more complex model, H_1 , was compared to the simpler model, H_0 , to provide a measure of significance for a better fit to the data based on additional parameters. After three million generations, model 6A was not significantly performing better than model 6B (Table 16). Similarly, model 7A could not be tested with model 7D at this number of generations because the logarithm of likelihood ratio between these two models was not within the percentage points of the χ^2 distribution (Table 16). After six million generations, model 6A improved marginally but still was not significantly better than

model 6B. However, model 7A improved greatly from three million to six million generations, becoming marginally better than model 7D with a P value close to .1 (Table 16). Although this indicates model 7D performed as good as model 7A, it suggests that model 7A improves over longer generations, while model 7D performed best between three and six million generations. The results warrant further analysis with extended generations to evaluate whether the trend over six million generations continues to move in favor of model 7A significantly outperforming model 7D by allowing for all thirty rate classes to evolve separately within the seven-state model of RNA evolution.

While the LRT provides a means to evaluate the different models within each class of RNA models, it does not allow for comparison of models from different classes, as six-state models are not nested within seven-state models. Furthermore, doubt has been cast on both the AIC and LRT as evaluators of models given that they rely on the data being asymptotic, a characteristic atypical of most phylogenetic datasets (Goldman, 1993). Alternatively, Cox's test (Cox, 1962) permits non-nested models to be compared; however, given the size of this dataset and the computation time required to perform Cox's test, it is not feasible for this study. Alternatively, I decided to evaluate the models based on the results provided here and objectively choose the models that appear to be performing well or improving over the increasing number of generations.

In determining which model class is better for this dataset, it is necessary to consider whether a mismatch parameter is necessary to include at all given the rarity of non-canonical base-pairs. The base-pair frequencies for best likelihoods of all six- and

seven-state models suggest that non-canonical basepairs occur in near equal frequency as UG intermediates (model 7D) and UG GU intermediates when double transitions are modeled separately (Table 17). Regardless of the seven-state model implemented, it appears that changes to and from non-Watson-Crick base-pairs occur nearly as frequently as GU UG intermediates. Excluding them will inflate rate estimates in six-state models, and will also exclude a significant proportion of the informative substitutions. Although the LRT did not indicate significantly a better fit to the data for model 6A over 6B, it did show marginal support for model 7A over 7D with increasing number of generations. Therefore, seven-state models may be preferred over six-state models. It seems biologically justified to include changes to and from the mismatch into maximum likelihood models of RNA evolution, as these substitutions represent a major component of the evolutionary processes in these molecules.

Table 16. Likelihood ratio tests performed within the six and seven-state models.

3 million							6 million			
H₀	H₁	d.f.¹	Best		Mean		Best		Mean	
			δ²	P(LRT)³	δ²	P(LRT)³	δ²	P(LRT)³	δ²	P(LRT)³
6B	6A	12	1.91	0.99>P>0.975	2.21	0.975>P>0.95	-2.29	NA	3.72	0.9>P>0.5
7D	7A	17	-8.01	NA	1.03	NA	10.47	0.5>P>0.1	9.69	0.5>P>0.1

¹ degrees for freedom = difference in the number of parameters between models
² δ = logarithm of the likelihood ratio: δ = ln(L₁/L₀), or ln L₁ - ln L₀
³ Note: according to χ² distribution of 2δ. NA = probability not available in percentage points χ²_{α,v} of the chi-squared distribution.

Table 17. Observed and model basepair frequencies within the helices of the 28S rRNA expansion segments and related core elements¹.

Basepair	Observed	(6-state)		Observed	(7-state)	
		6A	6B		7A	7D
AU	0.1572	0.2098	0.2235	0.1403	0.2169	0.2331
GU	0.0901	0.0765	0.1078	0.0748	0.0836	0.1023
GC	0.3607	0.2290	0.1954	0.3606	0.2393	0.1848
UA	0.1286	0.2028	0.2150	0.1152	0.1539	0.2136
UG	0.0931	0.0847	0.0634	0.0755	0.0534	0.0679
CG	0.1703	0.1972	0.1950	0.1636	0.1918	0.1345
MM	-----	-----	-----	0.0701	0.0611	0.0637

¹Calculated in PHASE ver. 1.1.

Model 7A or 7D?

Published phylogenetic studies using mixed RNA models in PHASE have all used model 7A (Hudelot *et al.*, 2002; Jow *et al.*, 2003; Gillespie *et al.*, 2005), and performance of models 7A and 7D have not been compared with empirical datasets. One way to evaluate how well a model is performing is to compare the number of clades recovered from independent analyses run at different starting seeds, with different generation times (Miller *et al.*, 2004). In this study, the two analyses performed under model 7A have many similarly recovered clades when superimposed upon one another (Fig. 18). Of particular interest is the same backbone structure of the trees, as it has previously been difficult to support this area of estimated phylogenies (Gillespie *et al.*, 2003, 2004a). Comparing the results of the two analyses performed under model 7D illustrates that fewer clades were recovered as compared to model 7A, with considerable differences in topology in the backbone of the superimposed trees (Fig. 19). While interesting, the results of these comparisons need to be interpreted with some caution. Firstly, since model 7A is steadily improving from three to six million generations, as suggested by mean and ESS over the two analyses (Table 5) and plots of likelihood against generations (Fig. 6), it could very well be that the sampled trees and tree lengths are already consistent after three million generations. Even with ten poorly sampled parameters, model 7A seems to improve with increasing significance compared to model 7D as suggested by the LRT (Table 10). However, it cannot be ruled out that model 7D performed better from three to six million generations, with sampled trees and tree lengths improving from three to six million generations. It is strange that, despite having



Figure 18. Superimposed branches of the trees generated under model 7A for three and six million sampling generations. Branches recovered in both analyses are colored red, with dissimilar clades colored black. Trees were generated using the *mcmcsummarize* program in PHASE ver. 2.0. Phylograms are an extended majority rule consensus of the three and six thousand sampled trees.

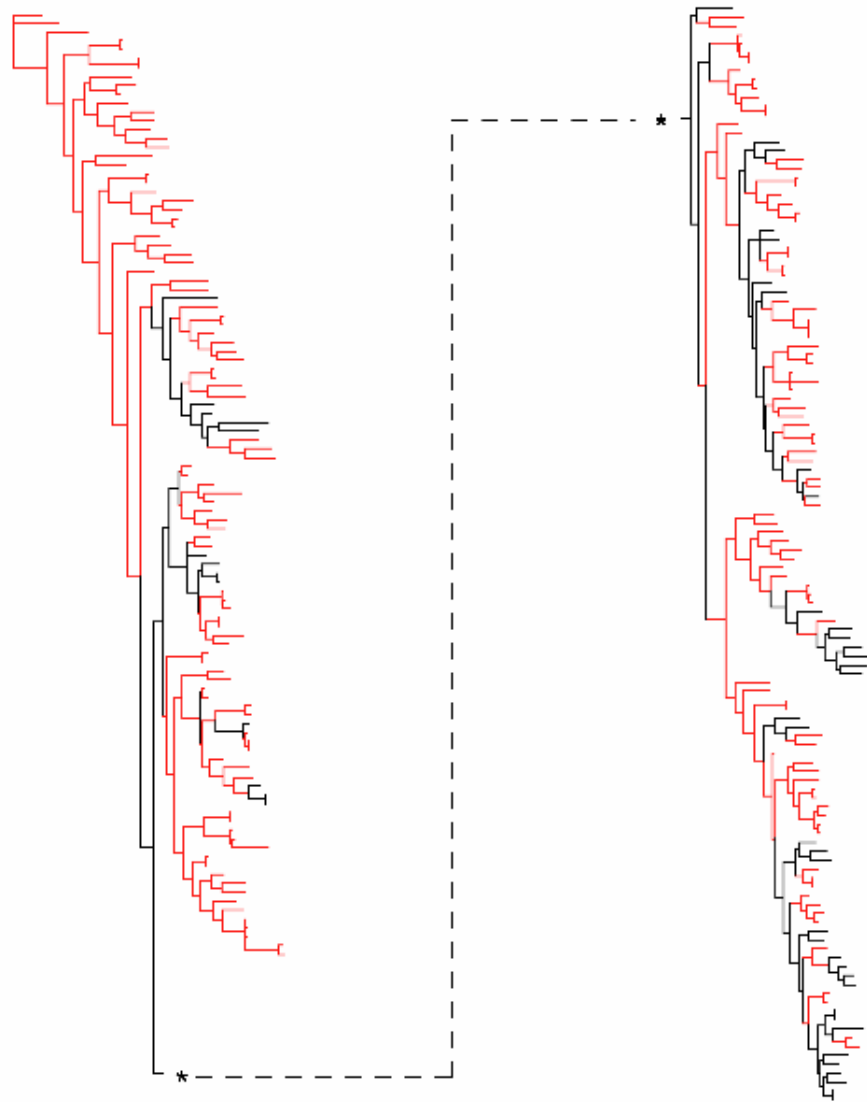


Figure 19. Superimposed branches of the trees generated under model 7D for three and six million sampling generations. Branches recovered in both analyses are colored red, with dissimilar clades colored black. Trees were generated using the *mcmcsummarize* program in PHASE ver. 2.0. Phylograms are an extended majority rule consensus of the three and six thousand sampled trees.

all of its parameters converge after three million generations, model 7D's likelihood still had a somewhat poor ESS even after six million generations. Regardless, it is tempting to select model 7A over 7D in light of its increase in significance as reported by the LRT. However, a preference would at best be premature, and not until the analyses are run for further generations (10 to 20 million) will the results of the LRT likely reach significant levels where one model can confidently be shown to perform as good as or better than the other. Until then I suggest that both models be considered in the phylogenetic analysis of RNA molecules.

Conclusion

Using the program PHASE, I compared six maximum likelihood models of RNA evolution that relax base-pair symmetry and permit non-zero rates for double substitutions across basepairs. Given the large number of taxa (231) and moderate number of characters (683), features typical of many phylogenetic datasets, 16-state models were intractable, requiring a large amount of computation, and only reaching completion in over three weeks, with many model parameters not reaching convergence. Six-state and seven-state models were compared over three and six million sampling generations via Markov Chain Monte Carlo simulation, providing model statistics to evaluate the performance of each model. Using the LRT, I determined that more general models (models 7A and 6A) were not significantly better than simpler models (models 7A and 7D), and that some parameters in these general models did not reach stationarity even after six million sampling generations. However, I should have used the LRT prior

to analysis by testing these models on the *same* tree (e.g. from parsimony or neighbor-joining), rather than comparing the average likelihoods obtained from *thousands of trees* that are not the same across all of the analyses. In this regard, the results of the LRT should not be given much weight. By estimating the basepair frequencies under all models I determined that non-canonical basepairs occur as frequently as GU and UG intermediates, suggesting that seven-state models more accurately model the substitution processes in RNA molecules. Furthermore, by estimating the frequency of substitution classes, I detected an asymmetry for double transitions, with CG→UA, UA→CG substitutions occurring at higher rates than AU→GC, GC→AU substitutions. The biological significance of this bias is unknown, but it suggests independent parameterization of double transitions should be implemented in phylogenetic studies (thus favoring models 7A and 6A over models 7D and 6B). Finally, the increasing significance of the model 7A outperforming model 7D after six million sampling generations, as reported by the LRT (again, interpret loosely), suggests that the more general seven-state model may outperform model 7D after further sampling generations are implemented. The results presented here will be of interest to those implementing RNA maximum likelihood models into phylogeny estimation. They are of particular interest because they provide the first empirical study that combined pairing and non-pairing models of evolution for a large number of taxa and a gene region that is commonly used for higher level phylogenetic reconstructions. Future work will determine the necessary analytical procedures to adequately implement these models such that they are useful for datasets of this nature.

Experimental procedures

Taxa examined

All sequences used in this study were generated from previous analyses on this beetle group (Gillespie *et al.*, 2003; Gillespie *et al.*, 2004a, b). Lists of the beetles analyzed here, with locality information, taxonomic position, and respective GenBank accession numbers for all sequences are provided in those studies. Voucher specimens for all sampled taxa have been deposited in the Texas A&M University, Rutgers University, or the University of Delaware insect museums. Information regarding sampled taxa is available at the following website (<http://hisl.tamu.edu>).

Multiple sequence alignment and scripted manipulation

The structural alignment and annotation of the rRNA dataset, along with methodology and relevant citations, is provided in Gillespie *et al.* (2004b) and is available at the jRNA website (<http://hymenoptera.tamu.edu/rna>) following the links to “Galerucinae-Alignments”. A secondary structure diagram of the 28S rRNA expansion segments D2 and D3 illustrates the helical, non-pairing, and excluded regions of the alignment (Fig. 11) and reflects the annotation provided in Gillespie *et al.* (2004b). The criterion for excluding regions is described in Gillespie (2004). I added a pairing mask (Hudelot *et al.*, 2002; Jow *et al.*, 2003) to the structural alignment that identifies each basepair within the rRNA molecule, as supported initially by covariation analysis and subsequently by the calculation of basepair frequency tables using the Jrna scripts available at the jRNA webpage. Additionally, I constructed a helix index file that

describes all pairing, non-pairing, and excluded regions of the alignment. Used in conjunction with the Jrna Perl scripts, the modified alignment and helix index produced the input file and control files necessary to analyze the six models within the PHASE program. Using the jRNA scripts I created two Nexus files separating paired and unpaired regions of the alignment. The file containing the unpaired regions of the alignment was analyzed in ModelTest (Posada & Crandall, 1998) to provide the best model of evolution for these positions in the alignment. The model of Tamura & Nei (TN93, 1993) was reported as the best model under the hierarchical likelihood ratio test (hLRT, Posada & Crandall, 1998, 2001) and second best to model HKY85 under the Akaike information criterion (AIC, Linhart & Zucchini, 1986). Given that TN93 distinguishes between different transition classes, and that TN93 is the optimal model under both the hLRT and AIC tests for 18S rRNA loops from the same beetle taxa (Gillespie, unpublished data from Chapter V), I elected to choose it over model HKY85. Thus all models analyzed in PHASE implemented the TN93+gamma+invariant sites model for non-pairing regions of the structural alignment.

Phylogeny estimation

For all six models, two control files were created that differed in their starting seed. Each model class contained two control files per model; one designed for a sampling of three million generations, the other designed for a sampling of six million generations. This resulted in 12 control files that all varied in their starting seeds, thus guaranteeing that each model could be compared with an independent analysis of twice the sampled

generation time. Like MrBayes, PHASE analyzes the data under maximum likelihood using Bayesian inference. I sampled every 1000 generations throughout each analysis using six Markov chains, keeping all chains at the same temperature and saving all branch lengths throughout. I used flat priors for all analyses presented here. All analyses were performed on Xblast (Texas A&M University), a 21 compute element (42 cpus) cluster of Apple G4 Xserves running iNquiry (<http://xblast.tamu.edu/>). Initial analyses were performed to determine the burn-in, or time until an acceptable plateau is reached in the sampling of likelihoods, trees and parameters in the posterior probability distribution. These burn-in values were determined by plotting log likelihoods ($-\ln L$) and tree lengths (TL) over generation number in the program Tracer ver. 1.2.1 (Rambaut & Drummond, 2005). Ultimately, a highly conservative value of 500,000 generations was selected for the burn-in prior to each of the twelve analyses.

Model evaluation

Results files from all analyses were modified with the Jrna scripts to produce input files for Tracer. To determine that both analyses per model reached a similar sampling space in the posterior distribution, analyses of both three and six million generations were compared in Tracer. The recovery of similar results for tree lengths and topologies, clade posterior probabilities and parameter posterior probabilities from these iterations is a good indicator that stationarity has been reached and that the Markov sampling procedure is effectively sampling these statistics throughout the estimated sample sizes (ESS) (Huelsenbeck *et al.*, 2002; Miller *et al.*, 2004). To evaluate the performance of

models within the three classes, I used the AIC, as defined by $AIC = -\ln L + \text{number of model free parameters}$. This calculation penalizes log likelihood scores from models with many parameters; hence it is a better comparison statistic than just simply comparing likelihood values between models. I also used the likelihood ratio test (LRT) to compare the two models within classes, with H_0 the simpler model nested within the more general model, H_I . The LRT calculates the logarithm of the likelihood ratio between the two models: $\delta = \ln (L_I/L_0)$. If H_0 is true, then 2δ will be distributed according to a χ^2 distribution with the number of degrees of freedom equal to the difference in the number of model parameters between H_0 and H_I (Linhart & Zucchini, 1986). As a measure of significance, a small P (the probability that 2δ from the χ^2 distribution is greater than the observed value of 2δ) indicates that H_I has a better fit to the data. Large P values indicate that the additional parameters in H_I are not significantly improving the fit given by H_0 .

CHAPTER V

PHYLOGENY OF ROOTWORMS AND RELATED GALERUCINE BEETLES (COLEOPTERA: CHRYSOMELIDAE) BASED ON THE ANALYSIS OF PARTIAL 28S AND 18S rRNA AND COI GENE SEQUENCES

Overview

The Galerucinae (Coleoptera: Chrysomelidae) *sensu stricto* (true galerucines) comprise a large assemblage of diverse phytophagous beetles containing over 5000 described species. Together with their sister taxon, the flea beetles, which differ from true galerucines by having the hind femora usually modified for jumping, the Galerucinae *sensu lato* comprises over 13000 described species and is the largest natural group within the Chrysomelidae. Unlike the flea beetles, for which robust hierarchical classification schemes have not been erected, an existing taxonomic structure exists for the true galerucines, based mostly on the work by John Wilcox. In the most recent taxonomic catalog of the Galerucinae *sensu stricto*, five tribes were established comprising 29 sections housing 488 genera (Seeno & Wilcox, 1982). The majority of the diversity within these tribes is found within the tribe Luperini, in which two genera, *Monolepta* and *Diabrotica*, are known to contain over 500 described species. In this chapter, I extend the work from previous phylogenetic studies of the Galerucinae by analyzing four amplicons from three genes representing 249 taxa, providing the largest phylogenetic analysis of this taxon to date. Using the two seven-state RNA models from

the previous chapter, I combine five maximum likelihood models (RNA +DNA for the rRNAs, three separate DNA models for the COI codon positions) for these partitions and analyze the data under likelihood using Bayesian inference. The results of these two analyses are compared with those from equally-weighted parsimony. Instead of choosing the results from one optimality criterion over another, either based on statistical support or philosophical predisposition, I elect to draw attention to the similar results produced by all three analyses, illustrating the power of multiple methods corroborating one another as support for phylogenetic estimation. In general, the results from all three analyses are consistent with previous molecular phylogenetic reconstructions for Galerucinae, except that increased taxon sampling for several groups, namely the tribes Hylaspini and Oidini, has improved the phylogenetic position of these taxa. As with previous analyses, under-sampled taxa, such as the Old World Metacyclini and all sections of the subtribe Luperina, continue to be unstable, with the few taxa representing these groups fluctuating in their positions based on the implemented optimality criterion. Nonetheless, I report here the most comprehensive phylogenetic estimation for the Galerucinae to date.

Introduction

The Galerucinae is one of the largest assemblages of leaf beetles (Chrysomelidae), with 1849 described species in 488 genera (Wilcox, 1965; Wilcox, 1972a, b; Seeno & Wilcox, 1982). Together with their highly divergent sister taxon, the flea beetles (> 560 genera and 8000 species; Seeno & Wilcox, 1982), galerucines pose the largest

agricultural threat of all chrysomelid beetles, often severely damaging important crop species (Metcalf, 1994). Adults feed on leaves and/or flower parts (including pollen), while larvae usually feed exclusively on roots or leaves (Jolivet & Hawkeswood, 1995; Riley *et al.*, 2002). Because of their subterranean habitat, the larvae of many galerucines have not yet been described, hence limiting our true understanding of definitive host relationships for many species (Gillespie *et al.*, 2004a).

Adult galerucines are characterized by an oval to oblong body, with the head visible from above and inserted into the prothorax without a neck-like constriction at the base (Riley *et al.*, 2002). The antennae of most species are shorter than the body, either filiform or clavate, composed of 11 articulated antennomeres (rarely 10), and narrowly separated from one another on the frons between the eyes (Reid, 1995a,b; Riley *et al.*, 2002). The pronotum is truncate or emarginate laterally, often with a lateral bead present (Riley *et al.*, 2002). The tarsi are pseudotetramerous with a 5-5-5 tarsal plan. Flea beetles usually have the hind femora swollen with an internal sclerotized extensor apodeme near the distal apex (Newman, 1835); but not always, as some transitional taxa (reviewed in Suzuki & Furth, 1992; Furth & Suzuki, 1994) lack these typical "jumping" organs. Females of galerucines and flea beetles have a distinct apodeme attached to sternite 8, and all testes are held together in a common membrane, usually compacted into a single sphere (Reid, 1995a, b). The larvae of both groups have from 0-1 stemmata, short femoral setae, short pretarsi, and a lobate paronychial appendix (Reid, 1995a, b). Larvae usually have broad mandibles with >3 teeth and a penicillus, and the labial palpi are 2-segmented (Reid, 1995a, b). Egg bursters, when present, are confined

to the meso- and/or metathorax (Reid, 1995a, b). Larvae are found externally feeding on plants, or in soil, roots, stems, or leaf mining. Further characters specific to flea beetles are reviewed in Riley *et al.* (2002).

Galerucinae sensu lato (= Trichostomata)

First introduced as the Tribe Galerucites, the taxon Galerucinae Latreille (1802) was proposed as an assemblage of all true galerucines and flea beetles. Two subsequent studies differed on the relationships between flea beetles and galerucines (reviewed in Lingafelter & Konstantinov, 1999). Newman (1835) acknowledged that true galerucines do not leap and have the antennal sockets placed closer together on the frons than flea beetles do. These observations led Newman (1835) to propose the taxon Alticinae with subfamilial rank to represent the flea beetles. Alternatively, Stephens (1839) grouped the flea beetles and other Galerucinae as Galerucidae (but with subfamilial rank), acknowledging that many of these beetles have the metafemora swollen, but not all.

These differing schools of thought have been carried down through the years, with many studies supporting both paradigms. Studies supporting the sister taxon relationship between monophyletic Alticinae and Galerucinae are based mostly on the initial conclusions of Newman (1835) (Redtenbacher, 1874; Jacoby, 1908; Heikertinger, 1912, 1924, 1941; Maulik, 1926; Winkler, 1929; Ogloblin, 1936; Heikertinger & Csiki, 1939, 1940), as well as more morphological characters (Gressit & Kimoto, 1963; Mohr, 1966; Scherer, 1969; Bechyné & Springlova de Bechyné, 1976; Medvedev, 1982; Lopatin, 1984; Gruev & Tomov, 1986; Doguet, 1994; Furth & Suzuki, 1994;

Konstantinov & Vandenberg, 1996) and molecular sequence data (Farrell, 1998). Other investigations have concluded that the flea beetles and true galerucines are not reciprocally monophyletic and that they should be grouped into one taxon, the Galerucinae (following Latreille [1802] and Stephens [1839]), with the flea beetles taking the same rank as the other tribes of the true galerucines (Allard, 1860, 1866; Chapuis, 1875; Horn, 1889; Böving & Craighead, 1931; Crowson, 1955; Lawrence & Britton, 1994; Reid, 1995a, b; Crowson & Crowson, 1996). This concept of grouping flea beetles and galerucines together is based on similarities shared between the larvae of both groups, as well as transitional forms occurring in the adults (Lingafelter & Konstantinov, 1999).

While solid evidence seems to support both views on the relationships between flea beetles and true galerucines, the problem of assigning an appropriate rank for the alticines has been highlighted in recent phylogenetic studies of these beetles. Using adult morphological characters from representatives of the main galerucine and alticine lineages, Lingafelter and Konstantinov (1999) reconstructed a phylogeny that placed the flea beetles as a monophyletic group nested within the galerucine subtribe Luperina (assuming that the placement of *Stenoluperus nipponensis*, a "problematic taxon" (Suzuki & Furth, 1992), is still within this subtribe). The authors concluded that the flea beetles should be given a subordinate rank within the Galerucinae, such as Alticini; however, their taxon sampling did not support or adequately test the monophyly of the 5 galerucine tribes (the Metacyclini were not sampled). A recent reanalysis by Kim *et al.* (2003) of Lingafelter and Konstantinov's (1999) morphological data, coupled with

partial DNA sequences from the EF1- α , 28S rRNA and COI genes, revealed a hypothesis more consistent with the paradigm of Latreille (1802) and Stephens (1839). A monophyletic core of true galerucines subtended by a paraphyletic assemblage of flea beetles was recovered with moderate support (Kim *et al.* 2003), a result highly consistent with the findings of Reid (1995a, b), which were based solely on morphological characters.

This same mode of alticine paraphyly encompassing a monophyletic core of strict galerucines was recovered by Duckett *et al.* (2004) through the combined reanalysis of Farrell's (1998) 18S rDNA data and Reid's (1995a, b, 2000) morphological characters. Gillespie *et al.* (2004a) also recovered this same relationship among true galerucines and flea beetles through the analysis of partial DNA sequences from the 28S rRNA and COI genes. A much larger analysis of these same gene regions from 137 taxa also generated this paraphyly under equally and differentially weighted parsimony; however, a maximum likelihood analysis recovered the flea beetles and strict galerucines as monophyletic sister taxa (Gillespie *et al.*, 2004a). The taxa causing the flea beetles to be paraphyletic with respect to the strict galerucines are similar in the above-mentioned studies, and a review of these potential *incertae sedis* taxa is provided by Duckett *et al.* (2004).

With the most recent phylogenetic analyses of flea beetles and galerucines seemingly converging on a common hypothesis of monophyly of the true galerucines within a paraphyletic flea beetle assemblage (Reid, 1995a, b; Duckett *et al.*, 2004; Gillespie *et al.*, 2003; Kim *et al.*, 2003; Gillespie *et al.* 2004a), some workers have

suggested the establishment of both groups as tribes (Alticini and Galerucini) within the single subfamily Galerucinae (Reid, 1995a, b, 2000; Duckett *et al.*, 2004), with the rankings of lesser groups within the strict galerucines adjusted downward (Duckett *et al.*, 2004). However, this approach may be premature given that 1. the existing phylogenies with robust taxon samplings for the true galerucines (Gillespie *et al.*, 2003, 2004a) are highly concordant with the taxonomic scheme presented by Seeno & Wilcox (1982), 2. there is not enough available data from all of the major lineages of flea beetles to reconstruct a robust phylogeny that tests their relationship relative to the true galerucines, 3. *incertae sedis* taxa and other problematic genera (Suzuki & Furth, 1992; Furth & Suzuki, 1994; Duckett *et al.*, 2004) should not be used to sink any taxonomic scheme until their systematic placement is better known (*and proven with multiple exemplar taxa*), and 4. a paraphyletic alticini, as recovered in several recent studies (Reid, 1995a, b; Duckett *et al.*, 2004; Gillespie *et al.*, 2003; Kim *et al.*, 2003; Gillespie *et al.* 2004a) would certainly add more confusion than stability within the existing taxonomic framework. Given these points, here I refer to the assemblage of flea beetles and true galerucines as the subfamily Galerucinae *sensu lato* (hereafter Galerucinae *s. l.*), with the Galerucinae *sensu stricto* (hereafter Galerucinae *s. s.* or true galerucines) containing only the five tribes listed by Seeno & Wilcox (1982). The terms alticine and flea beetle are used interchangeably, representing a rankless taxon distinct from the Galerucinae *s. s.*, as in Gillespie *et al.* (2003, 2004a).

Galerucinae sensu stricto

Subordinate delineations within *Galerucinae s. s.* were first proposed by Chapuis (1875) but did not stabilize until Weise's (1924) revision. Most galerucine workers followed Weise's (1924) system of eight tribes, but many considered these groupings to be inadequate (Wilcox, 1965). In the last taxonomic listing of the *Galerucinae s. s.*, Seeno & Wilcox (1982) restricted the subfamily to the following five tribes: Oidini (Chapuis, 1875), Galerucini (Latreille, 1802), Metacyclini (Chapuis, 1875), Hylaspini (Chapuis, 1875) replacing Sermylini as noted by Silfverberg (1990), and Luperini (Chapuis, 1875). These tribes and the characters supporting them are discussed below.

Oidini (Chapuis, 1875). The Oidini was first elevated to the rank of tribe within the *Galerucinae s. s.* by Weise (1923). Oidine galerucines are superficially quite distinct within the subfamily, being often quite large with the elytra extraordinarily convex, making them highly ovate in appearance. These beetles have the antennal insertions relatively low on the frons, bifid tarsal claws, the posterior margin of the last ventrite with a brief rectangular lobe (often inflexed), and an aedeagus without basal spurs. It is a relatively small group with 183 cataloged species (Wilcox, 1971) in 7 genera (Seeno & Wilcox, 1982). The range of the Oidini is exclusively Old World tropical. A comprehensive modern taxonomic treatment of the genera does not exist. No phylogenetic study has been performed on the Oidini to date, and recent higher-level studies have included only one species from the genus *Oides* to represent the tribe (Reid, 1995a, b, 2000; Farrell, 1998; Gillespie *et al.*, 2003a, b; Kim *et al.*, 2003; Duckett *et al.*, 2004; Gillespie *et al.*, 2004a).

Galerucini (Latreille, 1802). The Galerucini was first elevated to the rank of tribe within the Galerucinae s. s. by Laboissiere (1921). The Galerucini is the second largest galerucine tribe, with 1013 cataloged species (Wilcox, 1971) in 123 genera arranged in 5 uncharacterized sections (Seeno & Wilcox, 1982). Many of the beetles of this tribe are conspicuously pubescent, a characteristic that is rare to absent in other true Galerucinae tribes. These beetles have the antennal insertions relatively low on the frons, bifid tarsal claws (rarely simple), the posterior margin of the last ventrite is truncate to weakly emarginated with an adjoining semicircular depression of variable form and development, and the aedeagus with prominent basal spurs. The anterior and posterior tibiae lack spurs in most species (Wilcox, 1965). Larvae are above ground and feed on leaves (Wilcox, 1965). This tribe is cosmopolitan in distribution, with a few genera purportedly having species present in the Old and New World. A comprehensive modern taxonomic treatment of the world genera does not exist. No phylogenetic study has been performed on the Galerucini to date; however, two recent higher-level studies have sampled adequately within the tribe and suggest that it is a monophyletic taxon within a paraphyletic Metacyclini (Gillespie, 2001; Gillespie *et al.*, 2003, 2004a). Farrell (1998) included *Galerucella* sp. in his phytophaga phylogeny, hypothesizing it to be basal to the remaining sampled Galerucinae s.s. The inclusion of three galerucines (*Dircema cyanipenne*, *Caraguata pallida*, and *Erynephala punticollis*) to Farrell's (1998) dataset by Duckett *et al.* (2004) did not recover the tribe as monophyletic. The three Galerucini (*Monocesta coryli*, *Galeruca tanacetii*, and *Diorhabda persica*) in Lingafelter

and Konstantinov's (1999) phylogeny did not form a monophyletic group, although the reanalysis of this study with three molecular markers did (Kim *et al.*, 2003).

Metacyclini (Chapuis, 1875). The Metacyclini was first elevated to the rank of tribe within the Galerucinae s. s. by Leng (1920). There are 259 cataloged species (Wilcox, 1971) within 37 genera (Seeno & Wilcox, 1982). These beetles have the antennal insertions relatively low on the frons, appendiculate tarsal claws (rarely bifid), the posterior margin of the last ventrite truncate to weakly emarginated with adjoining area flattened, and the aedeagus with prominent basal spurs. The tibiae have spurs in most species (Wilcox, 1965). Larvae are unknown (Wilcox, 1965) although that of *Metacycla* is purportedly on leaves. The genera are grouped geographically by Seeno and Wilcox (1982) into an Old World group and a New World group. No phylogenetic study has been performed on the Metacyclini to date; however, two more recent higher-level studies have sampled adequately within the New World genera and suggest that this group is monophyletic and sister to the Galerucini (Gillespie, 2001; Gillespie *et al.*, 2003, 2004). Only one Old World metacycline has been sampled in these studies, *Palaeophylia* sp., and it consistently subtends the Galerucini, causing the Metacyclini to be a paraphyletic taxon. The studies of Farrell (1998) and Lingafelter and Konstantinov (1999) did not sample the Metacyclini. The inclusion of six metacyclines (*Palaeophylia* sp., *Masurius violaceipennis*, *Exora* sp., *Malacorhinus sericeus*, *Pyesia* sp., and *Chthoneis* sp.) to Farrell's (1998) dataset by Duckett *et al.* (2004) failed to recover the tribe as monophyletic.

Hylaspini (Chapuis, 1875). The Hylaspini was first elevated to the rank of tribe within the Galerucinae s. s. by Wilcox (1965) as Sermylini, a name later found to be invalid (Silfverberg, 1990). This tribe is comprised of 394 cataloged species in 49 genera arranged in six loosely characterized sections (Seeno & Wilcox, 1982). These beetles have the antennal insertions relatively low on the frons, appendiculate tarsal claws, the posterior margin of the last ventrite of male with a short evenly-rounded lobe, and an aedeagus without basal spurs. Males and females have terminal spurs on the middle and hind tibiae (Wilcox, 1965). Known larvae are on leaves (Wilcox, 1965). Although records exist for two hylaspine species in North America (*Sermylassa halensis* and *Agelastica alni*), these are most likely interceptions (Riley *et al.*, 2002), so the Hylaspini should be considered strictly an Old World taxon. No phylogenetic study has been performed on the Hylaspini to date, and most existing higher-level studies have under-sampled the tribe (Reid, 1995a, b, 2000; Farrell, 1998; Gillespie *et al.*, 2003; Kim *et al.*, 2003; Duckett *et al.*, 2004). Gillespie *et al.* (2004a) sampled 4 species from 3 sections (*Aplosonyx quadripustulatus*, *Sermylassa halensis*, *Agelasa nigriceps*, and *Agelastica coerulea*) but failed to recover the tribe as monophyletic under all optimality criteria used.

Luperini (Chapuis, 1875). The Luperini is the largest of the galerucine tribes, with 3953 cataloged species (Wilcox, 1972a, b) in 272 genera arranged in 18 sections within three subtribes (Seeno & Wilcox, 1982). The tribe is cosmopolitan, with a few genera occurring in both the New and Old World. The Luperini was first recognized as a tribe within the Galerucinae s. s. by Leng (1920). These beetles have the antennal

insertions relatively high on the frons, appendiculate or bifid tarsal claws, the posterior margin of the last ventrite of male with or without a truncate lobe, and an aedeagus without basal spurs (spur-like projections unlike those of Galerucini-Metacyclini are rarely present). Males and females usually have terminal spurs on the middle and hind tibiae (Wilcox, 1965). Females must often be keyed to genus before they can be assigned to subtribe (Blake, 1958; Wilcox, 1965).

Implied by the common name “rootworms” often applied to this group, all of the known larvae are subterranean in habitat, feeding on the roots of vascular plants. In addition to leaves, many adult luperines eat flower parts, particularly the reproductive structures (Jolivet, 1977, 1991; Neilsen, 1988). Pollen feeding, considered to be the primitive feeding condition to all chrysomelids (Crowson, 1960; Samuelson, 1994; Reid, 1995a, b), has been recorded more in the Aulacophorina and Diabroticina than in the Luperina (Crowson & Crowson, 1996). Species of *Aulacophora*, *Acalymma*, and *Diabrotica* include pollen in their diets, particularly from Cucurbitaceae (Samuelson, 1994). Thus, adult feeding and foraging behaviors among the Aulacophorina and Diabroticina are more similar when compared to those of the Luperina.

Wilcox (1965, 1972a, b) cataloged the Luperini as three subtribes: Aulacophorina, Diabroticina, and Luperina. These subtribes and characters supporting them are discussed below.

Aulacophorina (Chapuis, 1875). Aulacophorines are Luperini with a truncate lobe on the apical margin of last ventrite of male, and most included genera have bifid tarsal claws. These beetles have a transverse impression of variable form at the mid-

length of the pronotum. This is strictly an Old World group of 535 cataloged species (Wilcox, 1972a) in 36 genera arranged in two uncharacterized sections (Seeno & Wilcox, 1982). Most species are found in tropical Asia and Africa. The genera have never received a collective taxonomic treatment.

Diabroticina (Chapuis, 1875). Diabroticines are Luperini without a truncate lobe on the apical margin of the last ventrite of males, with males usually having apical spurs on only middle and hind tibiae, and females with apical spurs on all tibiae (Wilcox, 1965). This subtribe is strictly New World and its genera are cataloged in four reasonably distinct informal sections. With few exceptions, these sections can be characterized as follows (largely after Wilcox, 1965):

- 1) Diabroticites: genera with bifid tarsal claws, male with simple middle tibia and without spur on anterior tibia
- 2) Cerotomites: genera with appendiculate tarsal claws, male with simple middle tibia and without spur on anterior tibia
- 3) Phyllethrities: genera with appendiculate (rarely simple) tarsal claws, male with middle tibia with incision or emargination on inner margin before apex and without apical spur on anterior tibia
- 4) Trachyscelidites: genera with appendiculate tarsal claws, male with simple middle tibia and apical spur on anterior tibia

Luperina (Chapuis, 1875). Luperines are Luperini with a truncate lobe on the apical margin, with the apical spurs of all tibiae variable. This is the largest and most

complex of the luperine subtribes. It is difficult to characterize except that its members usually have appendiculate claws (very rarely bifid or simple) and usually the apical margin of the last male ventrite has some form of rectangular lobe (very rarely simple). There are 2425 cataloged species (Wilcox, 1972b) in 196 genera arranged in 12 informal sections (Seeno & Wilcox, 1982). Most of these informal sections have never been characterized and at least some are likely of dubious value. However, one of these sections, the Monoleptites, is relatively well-characterized and found world wide. This group was treated as a subtribe of the Luperini by Wilcox (1965) but later reclassified as a section within the Luperina (Seeno & Wilcox 1982). These beetles are morphologically distinct and form a easily recognizable group composed of many poorly delimited genera. Of the other informal sections of the Luperina, seven are mentioned later in the present work. These include: Adoxiites, Scelidites, Phyllobroticites, Ornithognathites, Exosomites, Luperites, and Megalognathites.

In this chapter, I extend the work from previous phylogenetic studies of the Galerucinae by analyzing four amplicons from three genes representing 249 taxa, providing the largest phylogenetic analysis of this taxon to date. It is my attempt to identify natural groups that are supported as monophyletic, such that the system of tribal, subtribal and sectional taxonomic delineations can be evaluated within a phylogenetic context. As with my previous work, it is also my goal to report on the continued progress of taxon sampling, as well as identify taxa that need to be more adequately sampled to help resolve certain areas of reconstructed phylogenies. This study evaluates

taxonomic delineations previously proposed for the Galerucinae (Wilcox, 1965, 1972a, b; Seeno & Wilcox).

Results and discussion

Sampled gene regions

A recent study provided a predicted structural model of the expansion segments D2 and D3 of the 28S rRNA from 231 chrysomelid taxa, with emphasis on the Galerucinae (Gillespie *et al.*, 2004b). For matters of completion, I provide here some structural information on the sampled COI and 18S rRNA gene regions. The 456 codons sequenced from the COI gene are compared to a predicted structure of the COI protein for Insecta (Lunt *et al.*, 1996) in an attempt to detect any irregularities in the sequenced chrysomelid taxa (Table 18). Aside from one codon insertion in the flea beetle *Aedmon morrisoni*, the read frame of this portion of the COI was not perturbed with the addition of new sequences. Additionally, I predicted the secondary structure of variable regions V4 and V7-9 of Domains II and II, respectively, of the nuclear small subunit rRNA (18S) using published models as a benchmark (Fig. 1). This consensus structure does not add any new features to existing structural predictions for this region of 18S rRNA; however, it does provide a template for the alignment of other chrysomelid taxa, and is likely informative for other closely related beetle families. Structural information from both the COI and 18S gene regions facilitated the assignment of positional homology in these sequences and provided subdivisions within partitions (stems, loops, codon

Table 18. Relative conservation of amino acid residues from the sampled region of the cytochrome oxidase I gene (COI).

amino acid ¹	codon position	Lunt <i>et al.</i> (1996) pos., residue ²	cell region ³	state P/NP ⁴	general comments
001	001-003	093, F	I1	NP	F in most; changes to I and P (syn)
002	004-006	094, P	I1	NP	conserved P
003	007-009	095, R	I1	P	R in most; changes to S (syn) and W (non-syn)
004	010-012	096, M	I1	NP	M,L in most; change to N (syn)
005	013-015	097, N	I1	P	conserved N
006	016-018	098, N	I1	P	conserved N
007	019-021	099, ---	M3	NP	M in most; changes to L (syn)
008	022-024	100, ---	M3	P	S in most; changes to T (syn) and G (non-syn)
009	025-027	101, F	M3	NP	F in most; changes to L (syn)
010	028-030	102, W	M3	NP	conserved W
011	031-033	103, ---	M3	NP	L in most; change to F (syn)
012	034-036	104, L	M3	NP	L in most; change to W (syn)
013	037-039	105, P	M3	NP	conserved P
014	040-042	106, P	M3	NP	conserved P
015	043-045	107, ---	M3	P	S in most; changes to A (non-syn)
016	046-048	108, L	M3	NP	L in most; change to I (syn)
017	049-051	109, ---	M3	----	S,T,N (P), I,F,L,M (NP)
018	052-054	110, ---	M3	NP	F,L in most; change to V (syn)
019	055-057	111, L	M3	NP	conserved L
020	058-060	112, ---	M3	NP	L,I,V; change to N (non-syn)
021	061-063	113, ---	M3	----	S,T (P), V,F,L,M (NP)
022	064-066	114, ---	M3	P	S in most; changes to G (non-syn)
023	067-069	115, ---	M3	P	S in most; changes to G (non-syn)
024	070-072	116, ---	M3	NP	M,I,V,L,A (NP); change to K (non-syn)
025	073-075	117, ---	M3	NP	M,I,V,L,A (NP); change to T (non-syn)
026	076-078	118, ---	E2	P	E in most; changes to N,D (syn)
027	079-081	119, ---	E2	P	S,N
028	082-084	120, ---	E2	NP	conserved G
029	085-087	121, ---	E2	NP	A,V

Table 18. Continued.

amino acid ¹	codon position	Lunt <i>et al.</i> (1996) pos., residue ² cell region ³		state P/NP ⁴	general comments
030	088-090	122, G	E2	NP	conserved G
031	091-093	123, T	E2	P	conserved T
032	094-096	124, G	E2	NP	conserved G
033	097-099	125, W	E2	NP	conserved W
034	100-102	126, T	E2	P	conserved T
035	103-105	127, V	E2	NP	conserved V
036	106-108	128, Y	E2	P	conserved Y
037	109-111	129, P	E2	NP	conserved P
038	112-114	130, P	E2	NP	conserved P
039	115-117	131, L	E2	NP	conserved L
040	118-120	132, ---	E2	P	S in most; changes to A (non-syn)
041	121-123	133, ---	E2	----	S,T (P), A,G (NP)
042	124-126	134, ---	E2	P	N in most; changes to T (syn)
043	127-129	135, ---	E2	----	S,T (P) outgroup only, I,A,V,L,M (NP)
044	130-132	136, ---	E2	----	S,T (P), A,F,G (NP)
045	133-135	137, H	E2	P	conserved H
046	136-138	138, ---	E2	----	S,N,E (P), G,M (NP)
047	139-141	139, ---	E2	NP	conserved G
048	142-144	140, ---	E2	----	S,N,T (P), APG (NP)
049	145-147	141, S	E2	P	S in most; change to A (non-syn)
050	148-150	142, V	E2	NP	conserved V
051	151-153	143, D	E2	P	conserved D
052	154-156	144, ---	E2	NP	L in most; changes to M,F (syn)
053	157-159	145, A	E2	NP	A in most; changes to T (non-syn)
054	160-162	146, I	E2	NP	conserved I
055	163-165	147, F	E2	NP	conserved F
056	166-168	148, S	E2	P	conserved S
057	169-171	149, L	E2	NP	L in most; change to F (syn)
058	172-174	150, H	M4	P	H in most; change to Y (syn)

Table 18. Continued.

amino acid ¹	codon position	Lunt <i>et al.</i> (1996) pos., residue ²	cell region ³	state P/NP ⁴	general comments
059	175-177	151, ---	M4	NP	L in most; change to M (syn)
060	178-180	152, ---	M4	NP	conserved A
061	181-183	153, G	M4	NP	conserved G
062	184-186	154, ---	M4	NP	conserved I
063	187-189	155, S	M4	P	conserved S
064	190-192	156, S	M4	P	conserved S
065	193-195	157, I	M4	NP	conserved I
066	196-198	158, ---	M4	NP	conserved L
067	199-201	159, G	M4	NP	conserved G
068	202-204	160, ---	M4	NP	conserved A
069	205-207	161, ---	M4	NP	I,V,M
070	208-210	162, N	M4	P	conserved N
071	211-213	163, ---	M4	NP	conserved F
072	214-216	164, ---	M4	NP	conserved I
073	217-219	165, ---	M4	P	S,T
074	220-222	166, T	M4	P	conserved T
075	223-225	167, ---	M4	----	T,D (P), I,V,M (NP)
076	226-228	168, ---	M4	NP	I,V,M,A,L; change to Y (non-syn)
077	229-231	169, ---	I2	P	conserved N
078	232-234	170, M	I2	NP	conserved M
079	235-237	171, ---	I2	P	H,Q,R,K
080	238-240	172, ---	I2	NP	P,I; changes to S (non-syn)
081	241-243	173, ---	I2	----	Y,Q,E,K,T,S,N (P), M,A,V,L,I (NP)
082	244-246	174, ---	I2	----	K,N,D,S (P), G (NP)
083	247-249	175, ---	I2	NP	M,L,I; changes to T (non-syn)
084	250-252	176, ---	I2	----	T,W,K,S,N,E,Q (P), F,I,L,M,A (NP)
085	253-255	177, ---	I2	NP	M,V,L,F,I,P (NP)
086	256-258	178, D	I2	P	D in most; changes to E (syn)
087	259-261	179, ---	I2	P	K,R,S,Q

Table 18. Continued.

amino acid ¹	codon position	Lunt <i>et al.</i> (1996)		state P/NP ⁴	general comments
		pos., residue ²	cell region ³		
088	262-264	180, ---	I2	NP	M,A,I,L; changes to T,S (non-syn)
089	265-267	181, ---	I2	NP	P in most; changes to S (non-syn)
090	268-270	182, L	I2	NP	conserved L
091	271-273	183, F	I2	NP	F in most; changes to L (syn)
092	274-276	184, ---	I2	NP	P in most; changes to V,I (syn), to S (non-syn)
093	277-279	185, W	I2	P	conserved W
094	280-282	186, S	I2	----	S (P), A (NP)
095	283-285	187, V	M5	NP	V in most; change to I (syn)
096	286-288	188, ---	M5	NP	I,V,M,L,F,A; changes to T (non-syn)
097	289-291	189, I	M5	NP	I in most; change to L (syn)
098	292-294	190, ---	M5	P	conserved T
099	295-297	191, A	M5	NP	A in most; changes to T (non-syn)
100	298-300	192, ---	M5	NP	I,V,L
101	301-303	193, L	M5	NP	conserved L
102	304-306	194, L	M5	NP	conserved L
103	307-309	195, ---	M5	NP	conserved L
104	310-312	196, ---	M5	NP	L in most; change to I (syn)
105	313-315	197, S	M5	P	conserved S
106	316-318	198, L	M5	NP	conserved L
107	319-321	199, P	M5	NP	conserved P
108	322-324	200, V	M5	NP	V in most; change to I (syn)
109	325-327	201, L	M5	NP	conserved L
110	328-330	202, A	M5	NP	conserved A
111	331-333	203, G	M5	NP	conserved G
112	334-336	204, A	M5	NP	conserved A
113	337-339	205, I	M5	NP	conserved I
114	340-342	206, T	E3	P	conserved T
115	343-345	207, M	E3	NP	conserved M
116	346-348	208, L	E3	NP	conserved L

Table 18. Continued.

amino acid ¹	codon position	Lunt <i>et al.</i> (1996) pos., residue ²	cell region ³	state P/NP ⁴	general comments
117	349-351	209, L	E3	NP	conserved L
118	352-354	210, ---	E3	P	conserved T
119	355-357	211, D	E3	P	conserved D
120	358-360	212, R	E3	P	conserved R
121	361-363	213, N	E3	P	conserved N
122	364-366	214, ---	E3	NP	L in most; changes to I (syn)
123	367-369	215, N	E3	P	N in most; change to S (syn)
124	370-372	216, T	E3	P	conserved T
125	373-375	217, S	E3	P	S,T
126	376-378	218, F	E3	NP	conserved F
127	379-381	219, F	E3	NP	conserved F
128	382-384	220, D	E3	P	conserved D
129	385-387	221, P	E3	NP	conserved P
130	388-390	222, ---	E3	----	S,T (P), A,V,I (NP)
131	391-393	223, G	E3	NP	G in most; changes to S,D (non-syn)
132	394-396	224, G	E3	NP	conserved G
133	397-399	225, G	E3	NP	conserved G
134	400-402	-----	E3	NP	G insertion in <i>A. morrisoni</i>
135	403-405	226, ---	E3	P	conserved D
136	406-408	227, P	E3	NP	conserved P
137	409-411	228, ---	E3	NP	conserved I
138	412-414	229, L	E3	NP	conserved L
139	415-417	230, Y	M6	P	Y in most; changes to F (non-syn)
140	418-420	231, Q	M6	P	conserved Q
141	421-423	232, H	M6	P	conserved H
142	424-426	233, L	M6	NP	L in most; change to S (non-syn)
143	427-429	234, F	M6	NP	conserved F
144	430-432	235, W	M6	P	conserved W
145	433-435	236, F	M6	NP	conserved F

Table 18. Continued.

amino acid ¹	codon position	Lunt <i>et al.</i> (1996) pos., residue ²	cell region ³	state P/NP ⁴	general comments
146	436-438	237, F	M6	NP	conserved F
147	439-441	238, G	M6	NP	conserved G
148	442-444	239, H	M6	P	H in most; changes to P (non-syn) deletion in <i>C. populi</i>
149	445-447	240, P	M6	NP	conserved P
150	448-450	241, E	M6	P	E in most; change to D (syn)
151	451-453	242, V	M6	NP	V in most; changes to M,P,G (syn)
152	454-456	243, Y	M6	P	Y in most; changeS to E,D (syn)
153	457-459	244, I	M6	NP	I in most; change to S (non-syn)
154	460-462	245, L	M6	NP	conserved L
155	463-465	246, I	M6	NP	I in most; change to V (syn)

¹ Sampled amino acids 1 to 155 correspond to residues 93 to 246 of the complete COI protein.
² Amino acids conserved across insects are listed, with variable residues depicted with a blank (---).
³ The positions of the residues in regards to the mitochondrial membrane are given with the following abbreviations (I1,2 = internal loops 1 and 2; E2,3 = external loops 2 and 3; M3,4,5 = membrane-spanning helices 3-5).
⁴ A consensus state of polar (P) or non-polar (NP) is provided where applicable.

positions). These subdivisions allowed for the assignment of specific maximum likelihood models for a better estimate of phylogeny using statistical inference.

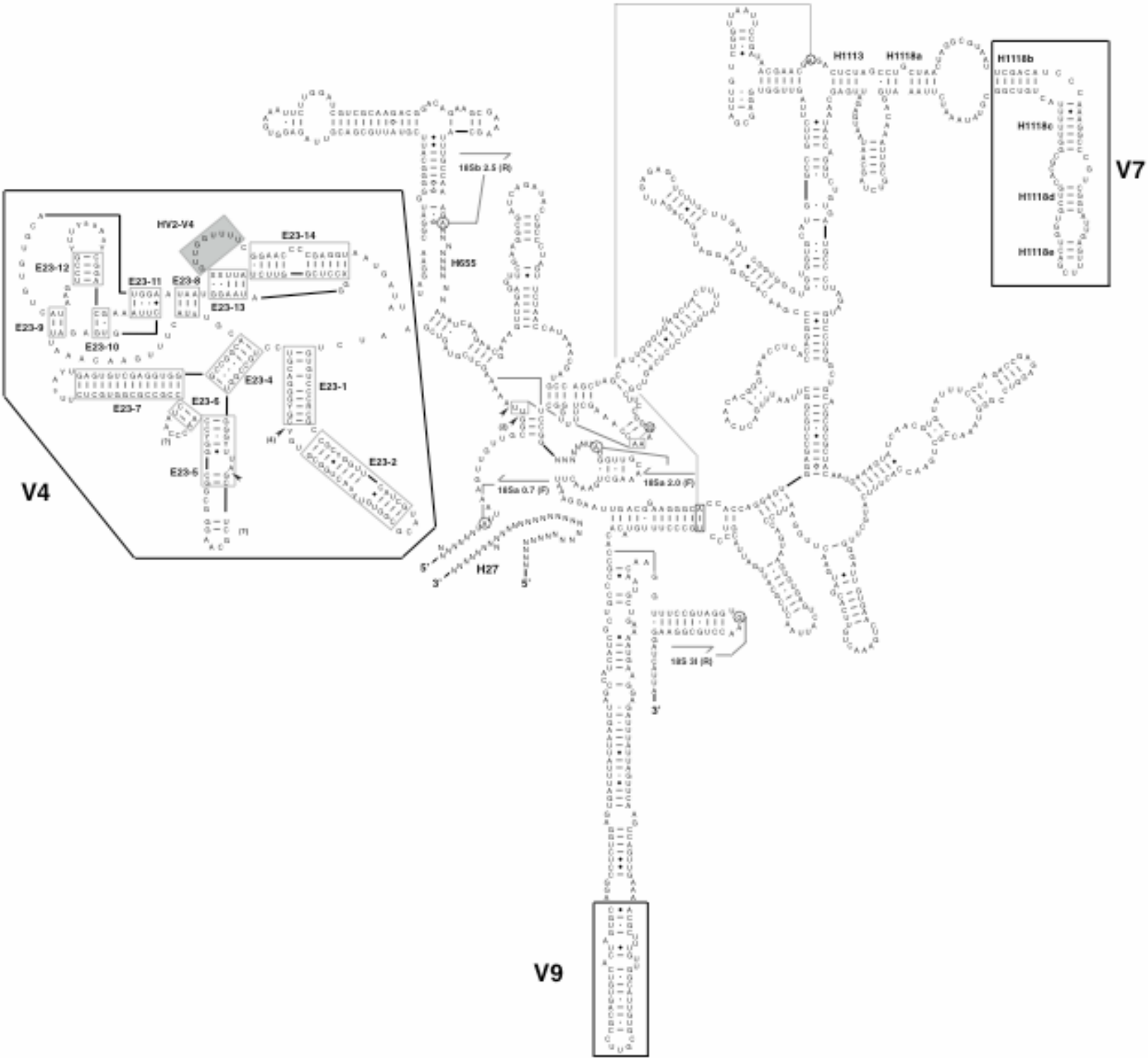
Computation, computation, computation...

The experimental design was ambitious and some of the proposed analyses could not be completed in a realistic timeframe (e.g., under one month). This is likely due to the large number of taxa and moderate number of characters sampled. Specifically, the Bayesian analyses using MrBayes, and all analyses using the program POY, did not complete within a month's time of analysis. Therefore, I am unable to report on the MrBayes and POY analyses. This, however, is not entirely disheartening for two reasons. First, even if the POY analyses did complete, it would be difficult to interpret the resulting trees compared to the other analyses, as none of the latter attempted to include information from unalignable regions. Second, a recent study showed that PHASE performs slightly better than MrBayes when implementing RNA models of substitution (Gillespie *et al.*, 2005), a result likely due to the more biologically-sound models offered in the PHASE program (see Chapter IV). Given this, I report only on the results of the Bayesian analyses run in PHASE and those obtained from equally weighted parsimony.

Equally weighted parsimony

Initial parsimony analyses were performed with the program PAUP* ver. 4.0b10 (Swofford, 2001); however, given the islandic nature of this large dataset, the shortest

Figure 20. Consensus secondary structure diagram of domains II and III of the nuclear small subunit rRNA (18S) gene region for the chrysomelids sequenced here. Conserved helices are within thin boxes, while variable regions are within thick boxes. Base-pairing (where there is strong comparative support) and base triples are shown connected by continuous lines. Base-pairing is indicated as follows: standard canonical pairs by lines (C-G, G-C, A-U, U-A); wobble G·U pairs by dots (G·U); A·G pairs by open circles (A●G); other non-canonical pairs by filled circles (e.g. C●A). Universal primers, as well as primers designed in this study, are mapped on the structure with the first primer position circled. A primer table is posted at the jRNA website. Diagram was generated using the program XRNA (Weiser, B. & Noller, H., University of California at Santa Cruz) with severe manual adjustment.



trees (in the thousands) found by PAUP* had a length of 10321. Although the strict consensus trees were very resolved, I decided to compare these results with a parsimony analysis using the program TNT (Goloboff, 1999; Goloboff *et al.*, 2003). This search strategy again yielded thousands of equally parsimonious trees, however, the length was much shorter than that obtained from PAUP (10308 steps). The consensus tree of the TNT analysis is shown in Figure 21.

Despite little support for the backbone of the tree, a problem encountered in previous studies (Gillespie *et al.*, 2003, 2004a), several natural groups are supported as monophyletic. As in these previous studies, the lupinine *Stenoluperus nipponensis* is grouped within the flea beetles. This is not surprising, as Furth and Suzuki (1992) have included *S. nipponensis* in their group of "problematic intermediates" between flea beetles and true galerucines. I do not consider the placement of *S. nipponensis* as evidence for a polyphyletic Galerucinae *sensu lato* (=Trichostomata), but rather continued support for this taxon as a flea beetle. Similarly, as with other studies, the Old World metacycline, *Palaeophylia* sp., is grouped within the tribe Galerucini. There are two interpretations for this result. First, *Palaeophylia* sp. is the only sampled Old World metacycline and under-sampling of these metacyclines could be forcing *Palaeophylia* sp. to not fall within the New World Metacyclini clade (clade 7 in Fig. 21, panel B). Alternatively, *Palaeophylia* sp. may actually be a galerucine and needs to be removed from the Metacyclini. Until more Old World metacyclines are included in such studies, this result should be treated with caution. Regarding the Metacyclini, a second result from equally-weighted parsimony is unacceptable from a morphological viewpoint. The

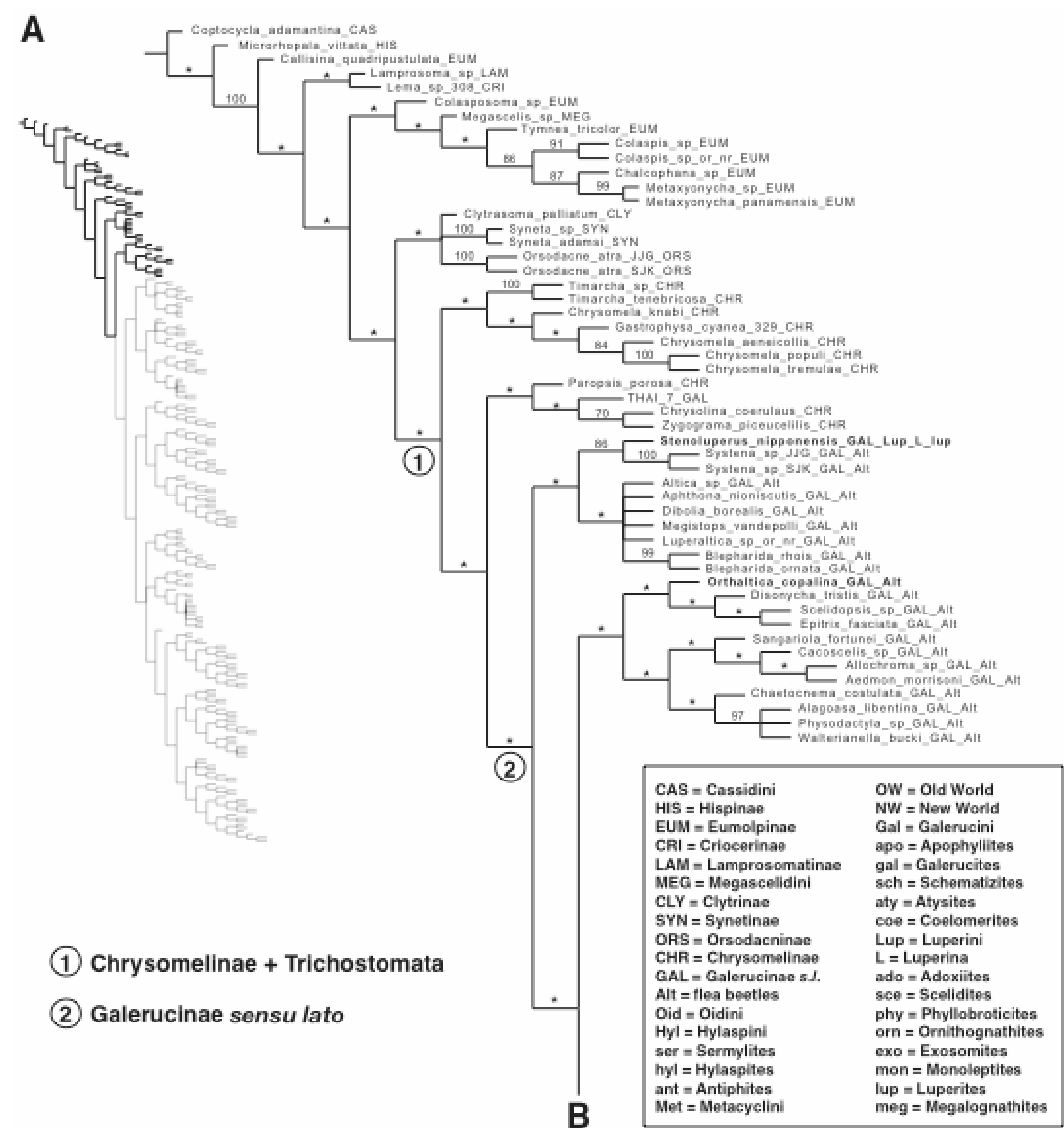


Figure 21. Results of an equally-weighted parsimony analysis of the combined data (COI nucleotides, 18S and 28S rRNA nucleotides). The tree is a strict consensus of 10,000 equally parsimonious trees of 10,308 steps. Branch support is from a 100 replicate bootstrap analysis with a 50 percent cut-off. Internodes with an asterisk were not recovered in the bootstrap analysis. Monophyletic groups are numbered one to 15 and are discussed in the text. Each taxon name is appended with one to several mnemonics. These mnemonics are explained in the taxon list enclosed within a box in panel A. The entire cladogram is minimized at left, with the portion that is enlarged in each panel bold. Taxa referred to specifically in the text are colored bold.

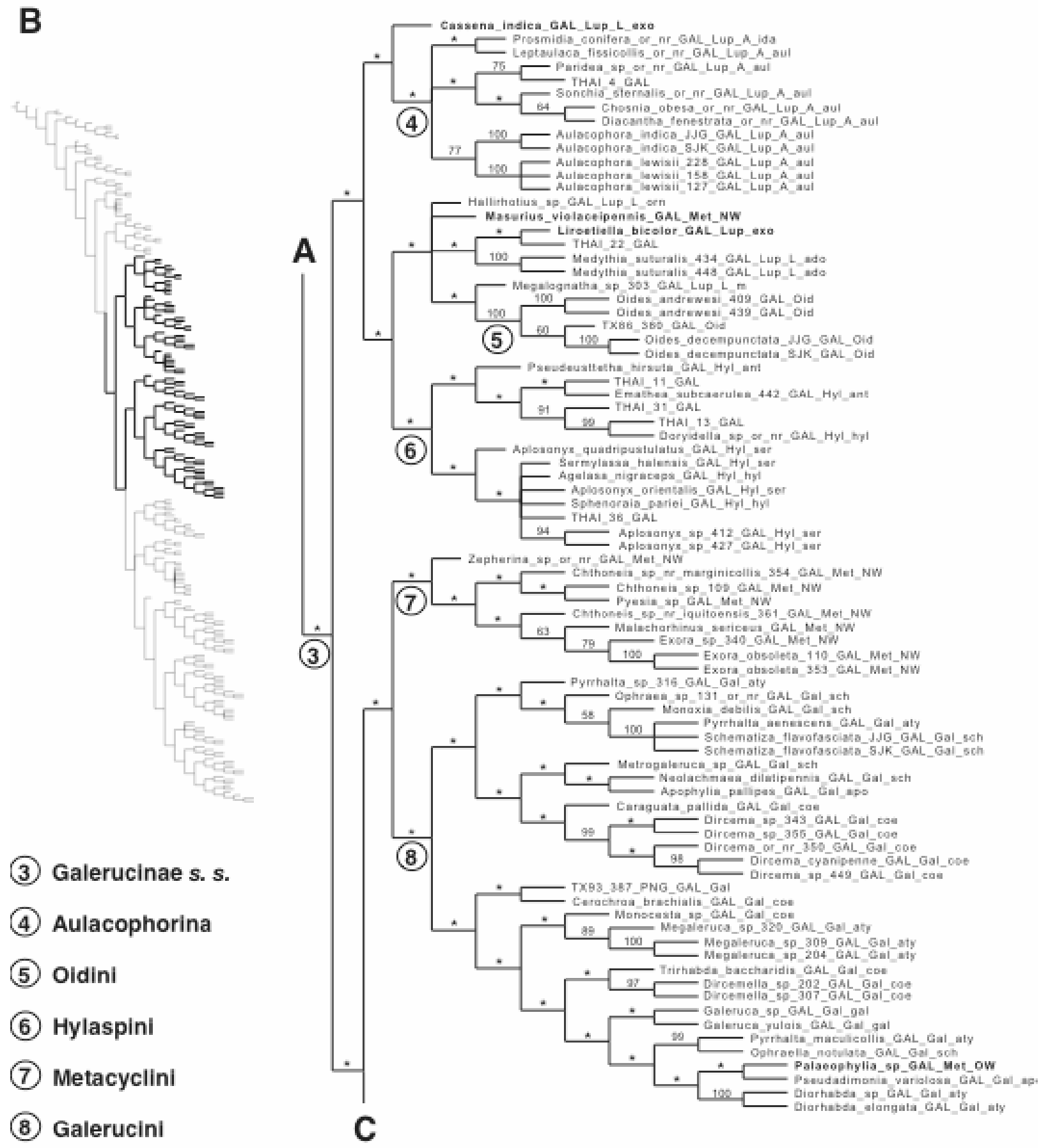


Figure 21 Continued.

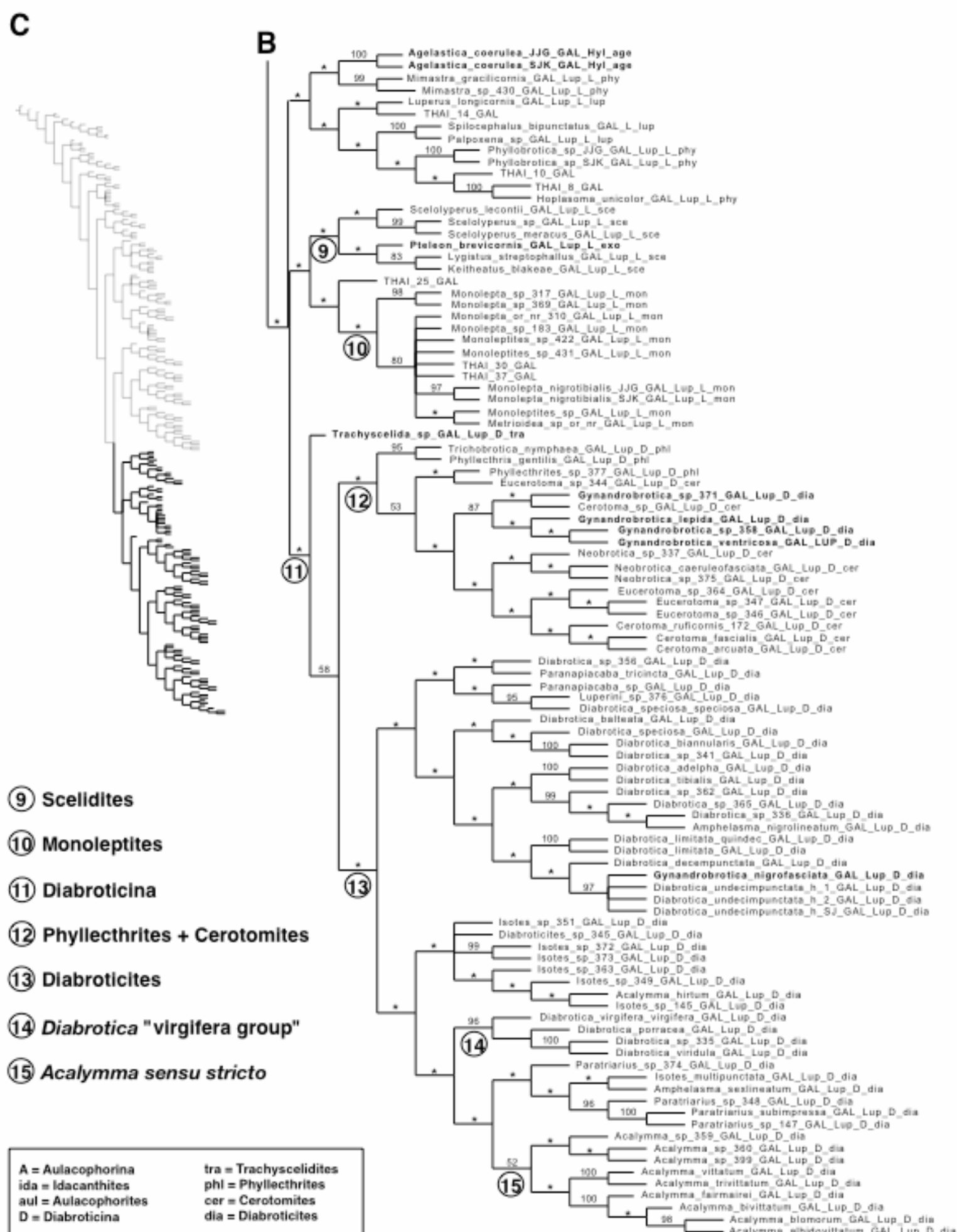


Figure 21 Continued.

placement of *Masurius violaceipennis* within a clade containing luperines and the tribe Oidini is surprising, as this taxon has always grouped with the Metacyclini in previous studies (Gillespie *et al.*, 2003, 2004a).

A very interesting result is the recovery of all four diabroticine sections (Trachyscelidites, Phyllethrines, Cerotomites, and Diabroticites) within a monophyletic Diabroticina (clade 11 in Fig. 21, panel C). Gillespie *et al.* (2004a) were unable to place a single species of *Trachyscelida*, the only representative of the monotypic section Trachyscelidites, within Diabroticina using equally weighted parsimony, differentially weighted parsimony, or modeling. Within the Diabroticites, *Gynandrobrotica nigrofasciata* continues to group close to *Diabrotica undecempunctata howardi* (southern corn rootworm), a result that is highly unlikely given that the other four species of *Gynandrobrotica* group within the Cerotomites. Because *Gynandrobrotica* is placed within the Diabroticites by Seeno and Wilcox (1982), previous studies did not elaborate on the polyphyly of *Gynandrobrotica*. However, evidence in this dissertation (Chapter IV) suggests that when analyzed using only the 28S rRNA, all five sampled species of *Gynandrobrotica* form a monophyletic group within the Cerotomites. Given this, I suspect that the COI sequence of *G. nigrofasciata* is a contaminant, and is in fact that of *D. undecempunctata howardi*. Since the 28S rRNA sequence of *G. nigrofasciata* appears to be valid, apparently the source of contamination was post-PCR; thus, the extraction code linking the DNA to the voucher specimen should remain intact. In light of this discovery I have flagged ascension number AY242451 on GenBank as a contaminant.

Another noteworthy result under parsimony is the placement of the exosomite *Pteleon brevicornis* within a monophyletic Scelidites (clade 9 in Fig. 21, panel C). This is not entirely surprising, as the other two sampled exosomites, *Liroetiella bicolor* and *Cassena indica*, do not group together and are not well supported in their placements within the cladogram. Either more exosomite taxa need to be collected to resolve this matter, or the Exosomites are not a natural group.

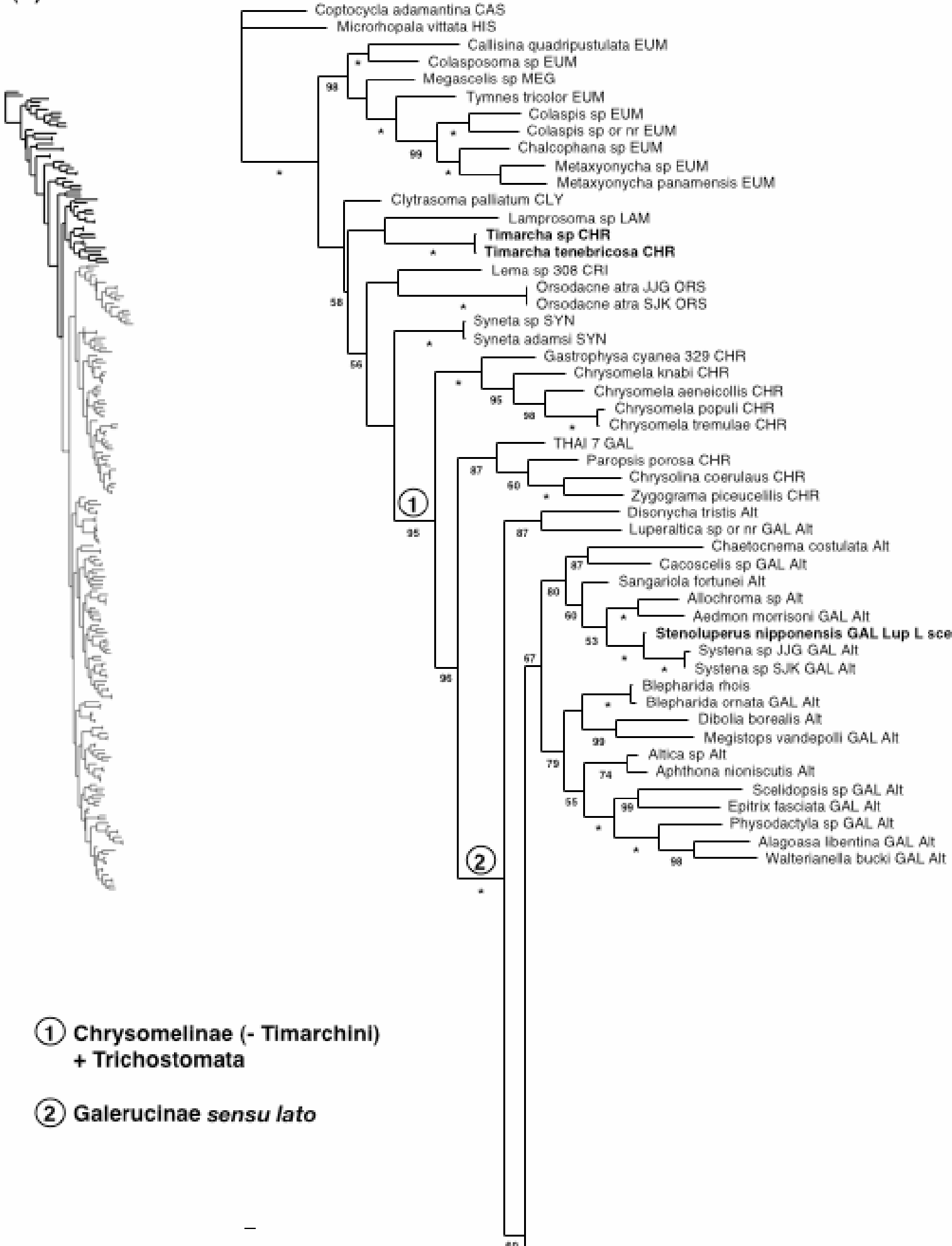
Finally, as with previous studies (Gillespie *et al.*, 2003, 2004a), the hylaspine taxon *Agelastica coerulea* does not group within the remaining tribe Hylaspini (clade 6 in Fig. 21, panel B). Along with a three-fold increase in the taxon sampling of the hylaspines, I included a second representative of *A. coerulea* in this study in an attempt to remedy this problem. Given that this second individual was sequenced in a separate lab (Kjer laboratory, Rutgers University, New Jersey), and it groups strongly with the original taxon, I conclude that *A. coerulea* is indeed difficult to place within the Hylaspini. From a morphological viewpoint *A. coerulea* should group with the remaining hylaspines.

Maximum likelihood (RNA7A)

The results of the combined five model maximum likelihood analysis using model 7A on the rRNA basepairs is shown in Figure 22. A curious result is found in the outgroup, where one of the two chrysomeline tribes, Timarchini, is not grouped with the remaining subfamily Chrysomelinae. This result is hard to interpret, but is perhaps the result of the outgroup sampling effecting the polarity of the Chrysomelinae. A second result

Figure 22. Extended majority rule consensus from a Bayesian analysis of the combined data (COI nucleotides, 18S and 28S rRNA nucleotides, 18S and 28S rRNA basepairs) under five maximum likelihood models (one per each partition). The rRNA basepairs were modeled under model 7A (see text for other partition models). Branch support values represent estimates of posterior probability. Internodes with an asterisk depict branches recovered with 100 percent posterior probability. Values below 50 percent are not shown. Monophyletic groups are numbered one to 15 and are discussed in the text. Each taxon name is appended with one to several mnemonics. See Figure 21 for a description of these mnemonics. The entire phylogram is minimized at left, with the portion that is enlarged in each panel bolded. Taxa referred to specifically in the text are colored bold.

(A)



(B) *

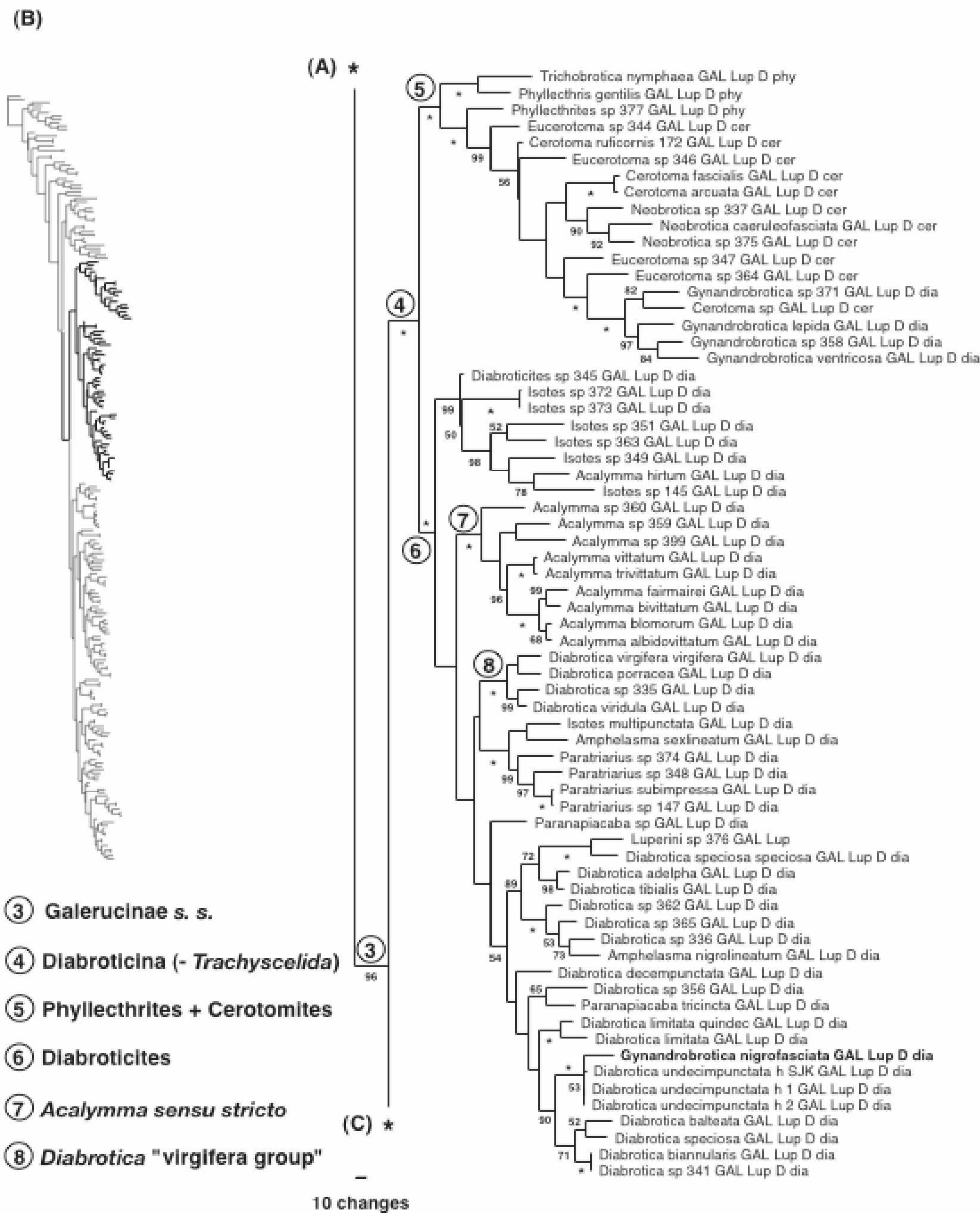


Figure 22 Continued.

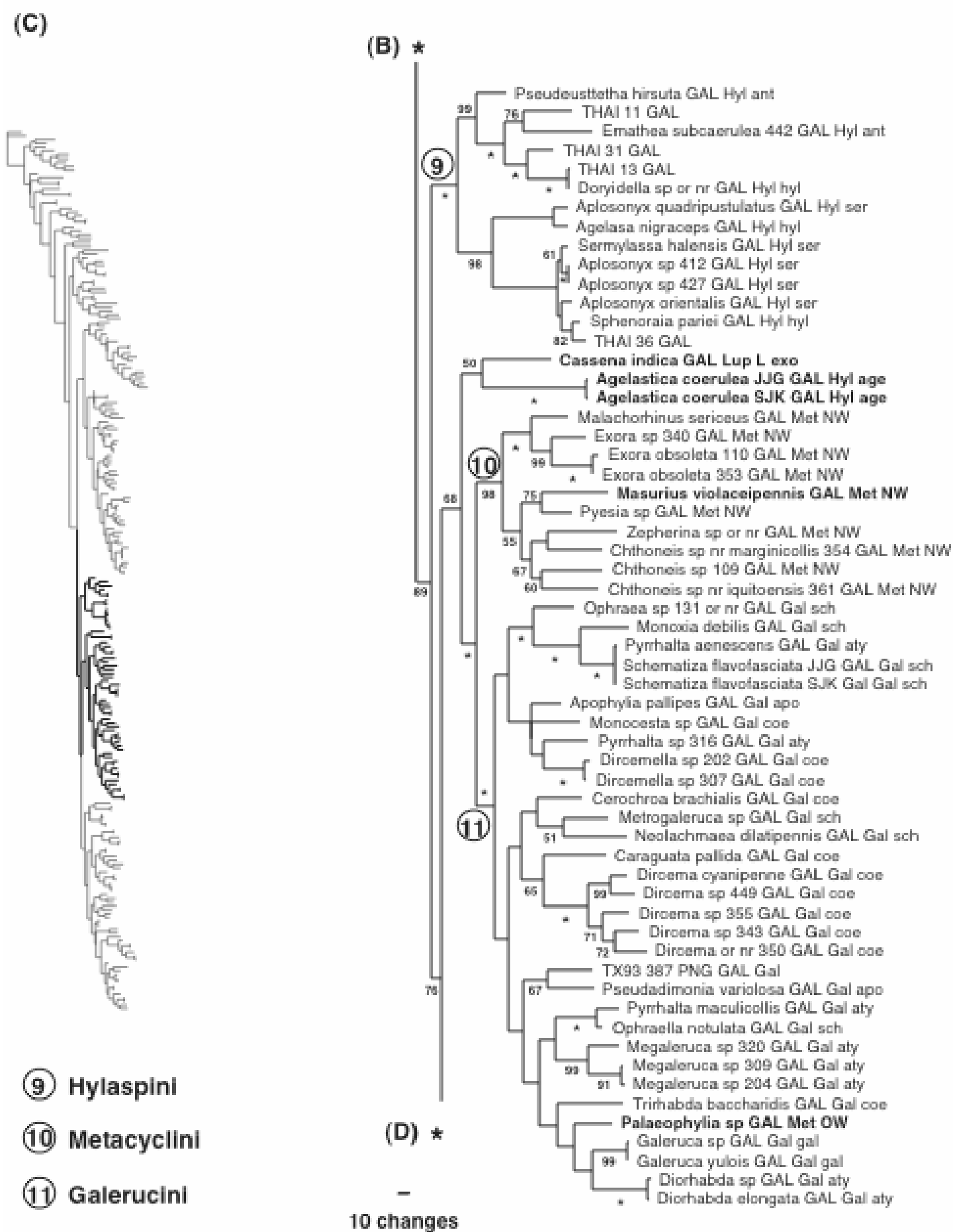


Figure 22 Continued.

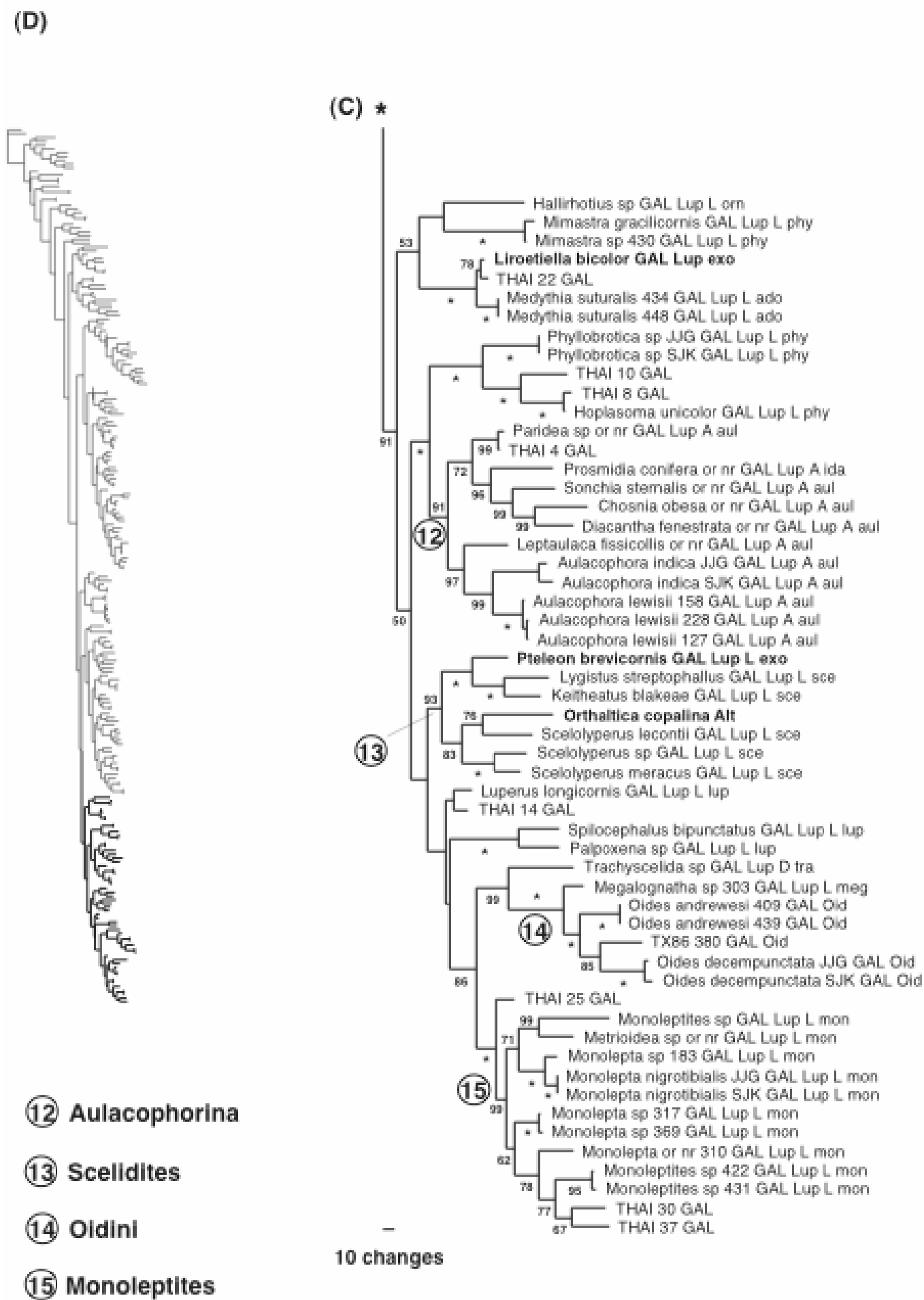


Figure 22 Continued.

concerning the outgroup is the placement of one flea beetle, *Orthaltica copalina*, in the middle of the monophyletic luperine section Scelidites (group 13 in Fig. 22, panel D). As with the placement of *S. nipponensis* in the parsimony analysis, this is not surprising, as Furth and Suzuki (1992) have included *O. copalina* in their group of "problematic intermediates" between flea beetles and true galerucines. However, because parsimony and the second likelihood analysis performed here (see below) did not recover *O. copalina* outside of the flea beetle group, I attribute the placement of this taxon within the Scelidites as an artifact of the analysis. As in parsimony (Fig. 21, panel A) and previous studies (Gillespie *et al.*, 2003, 2004a), the placement of *S. nipponensis* is within the flea beetle group (Fig. 22, panel A).

Unlike parsimony, this likelihood analysis did not recover *Trachyscelida* sp. within the Diabroticina, but instead nested in a group containing *Megalognatha* sp. and the Oidini (Fig. 22, panel D). Interestingly, this placement of *Trachyscelida* sp. near Oidini was recovered in the equally- and differentially-weighted parsimony analyses of Gillespie *et al.* (2004a), but not in the likelihood analysis. While *Trachyscelida* sp. will likely always be difficult to place within a phylogeny estimation, due to it being a monotypic section and the lack of sampled taxa, I conclude that parsimony probably revealed its proper position within the Galerucinae, assuming the inclusion of it within the Diabroticina by Seeo and Wilcox (1982) was based on some morphological evidence.

As in previous studies (Gillespie *et al.*, 2003, 2004a) the placement of the metacycline *M. violaceipennis* is within the Metacyclini (Fig. 22, panel C). The

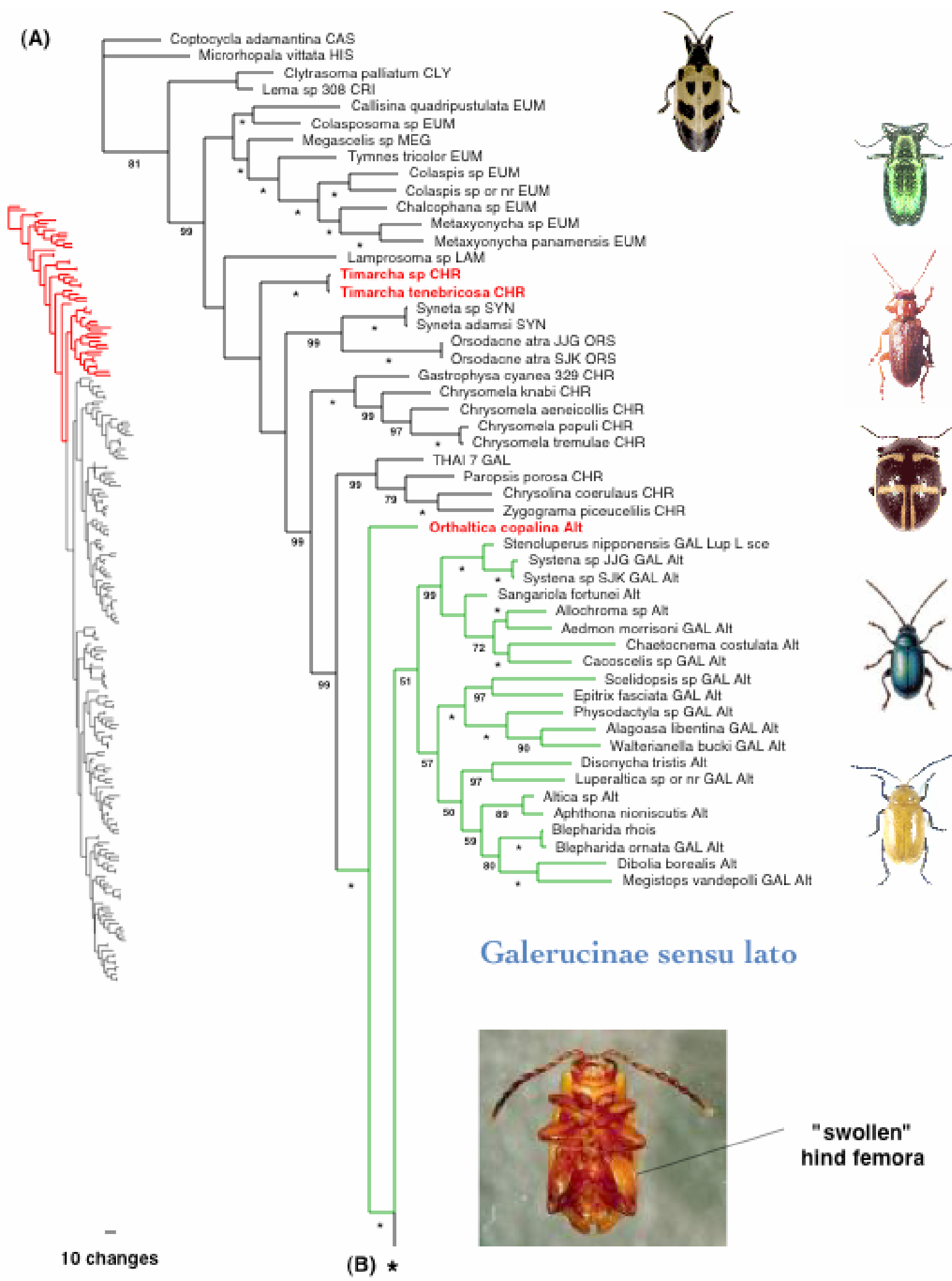
positions of the other problematic galerucine taxa described above, namely *A. coerulea*, *Palaeophylia* sp., *G. nigrofasciata*, *P. brevicornis*, are in agreement with the parsimony analysis. The two unstable exosomite taxa, *L. bicolor* and *C. indica*, are not sister groups or even adjacent and argue against a monophyletic Exosomites.

Maximum likelihood (RNA7D)

The results of the combined five model maximum likelihood analysis using model 7D on the rRNA basepairs is shown in Figure 23. The curious placement of *O. copalina* in the 7A maximum likelihood analysis was not recovered using model 7D, with the placement of *O. copalina* subtending the remaining flea beetles (Fig. 23, panel A). As with the 7A maximum likelihood analysis the Timarchini is split from the remaining Chrysomelinae. Also, in a agreement with the above analyses, the likelihood analysis using model 7D recovered *S. nipponensis* within the flea beetle group (Fig. 23, panel A).

Like the likelihood analysis using model 7A, this likelihood analysis did not recover *Trachyscelida* sp. within the Diabroticina, but instead nested in a group containing *Megalognatha* sp. and the Oidini (Fig. 23, panel B). However, this group immediately subtends the remaining three sections of the Diabroticina, a result recovered, interesting enough, by the equally-weighted parsimony analysis of Gillespie *et al.* (2004a). The positions of the other problematic galerucine taxa described above, namely *A. coerulea*, *Palaeophylia* sp., *G. nigrofasciata*, *P. brevicornis*, are in agreement with the 7A likelihood analysis and parsimony analysis. The two fluctuating exosomite

Figure 23. Extended majority rule consensus from a Bayesian analysis of the combined data (COI nucleotides, 18S and 28S rRNA nucleotides, 18S and 28S rRNA basepairs) under five maximum likelihood models (one per each partition). The rRNA basepairs were modeled under model 7D (see text for other partition models). Branch support values represent estimates of posterior probability. Internodes with an asterisk depict branches recovered with 100 percent posterior probability. Values below 50 percent are not shown. Monophyletic groups are defined as follows: O = Oidini, D = Diabroticina (- *Trachyscelida* sp.), p = Phyllecthrites, c = Cerotomites, d = Diabroticites, H = Hylaspini (- *Agelastica coerulea*), A = Aulacophorina, s = Scelidites, m = Monoleptites, M = Metacyclini (- *Palaeophyllia* sp.), G = Galerucini. Each taxon name is appended with one to several mnemonics. See Figure 21 for a description of these mnemonics. The entire phylogram is minimized at left, with the portion that is enlarged in each panel colored red. Taxa referred to specifically in the text are colored red. The *Trachyscelida* sp. + *Megalognatha* sp. + Oidini group referred to in the text is colored blue. The flea beetle group is colored light green. Species known to specialize on cucurbitacins are colored green. Pictures in descending order across the panels: Panel A: Hispinae, Eumolpinae, Orsodacnidae, Chrysomelinae, flea beetle 1, flea beetle 2; Panel B: Oidini, *Acalymma sensu stricto*, Diabrotica "virgifera group", *Diabrotica undecempunctata howardi* (southern corn rootworm); Panel C: Hylaspini, Aulacophorina, Scelidites, Monoleptites, Metacyclini, Galerucini 1, Galerucini 2, Galerucini 3.



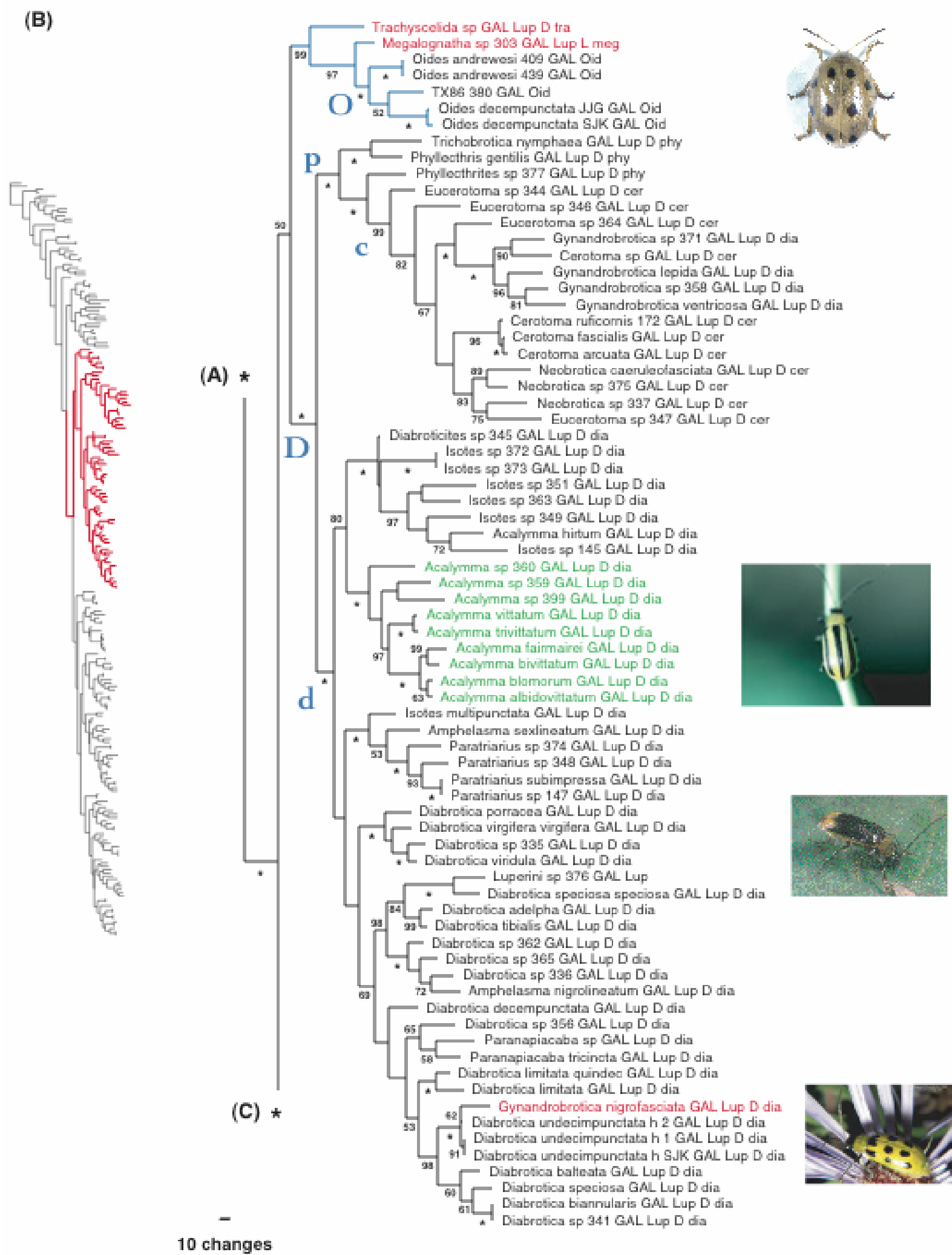


Figure 23 Continued..

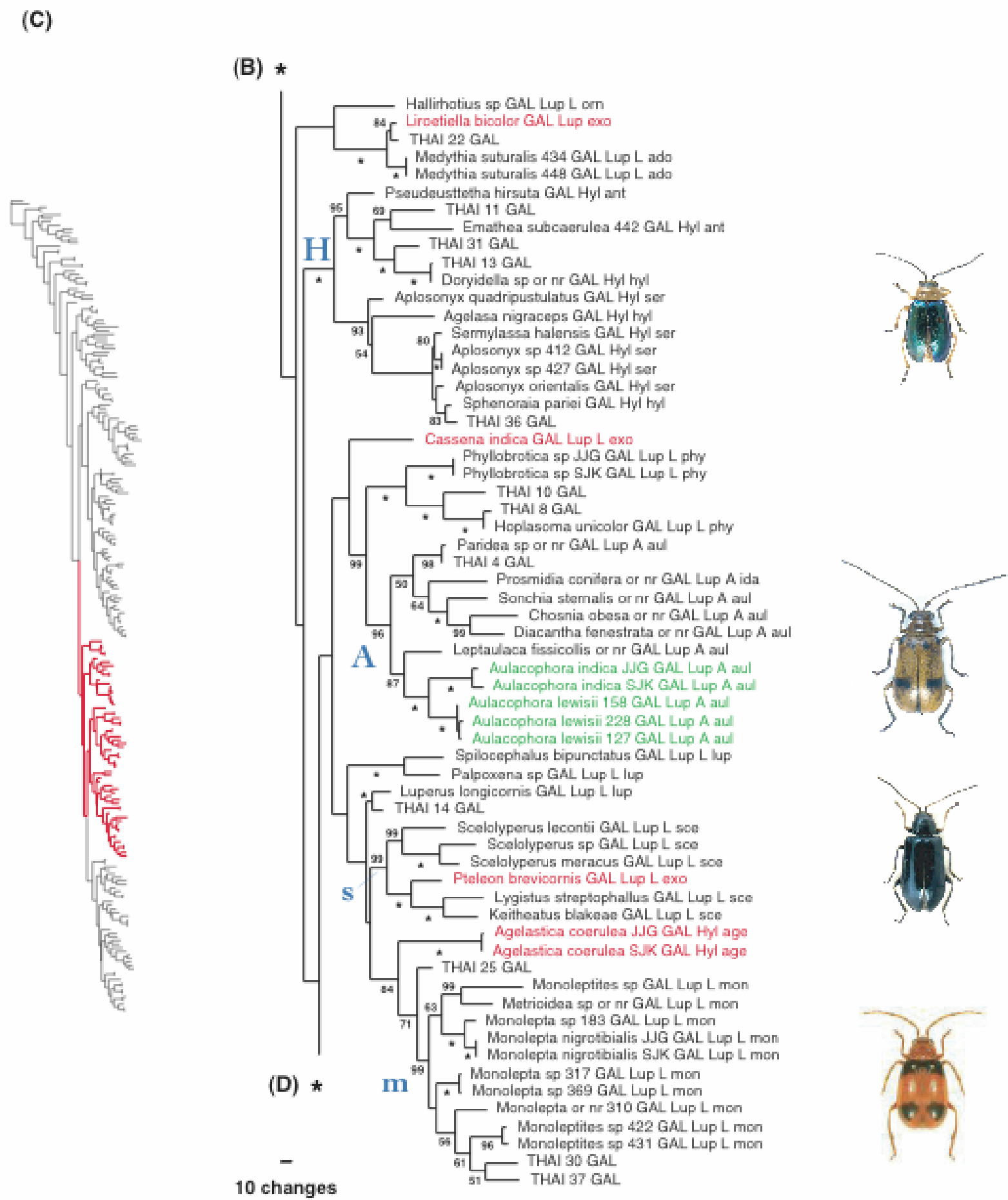


Figure 23 Continued.

(D)

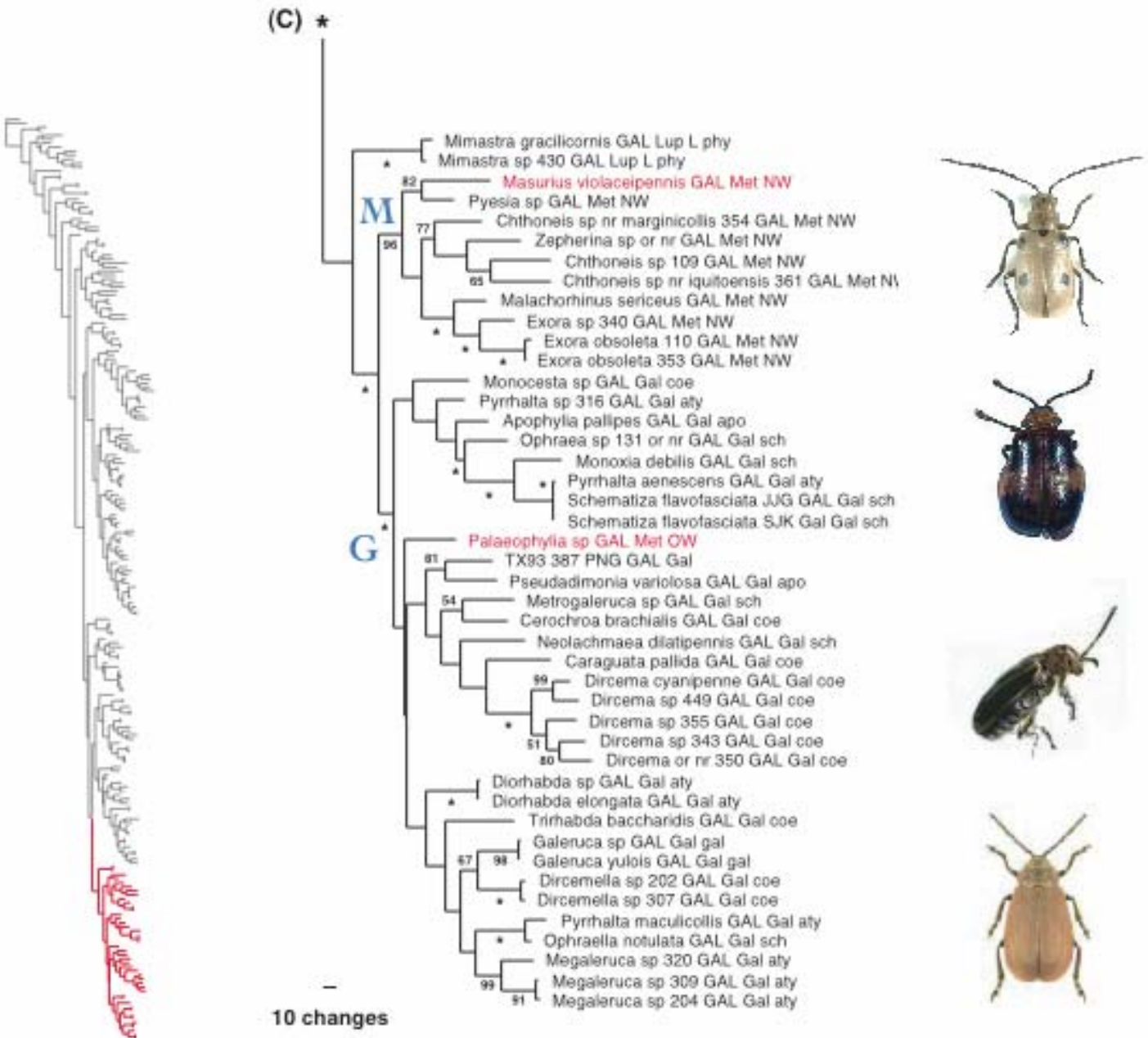


Fig. 23. Continued.

taxa, *L. bicolor* and *C. indica*, are not placed near one another and provide further support against a monophyletic Exosomites.

MCMC simulation

The maximum likelihood analyses performed under Bayesian inference were heavily parameterized, incorporating five models of substitution of the sub-partitions of the combined data. All model parameters were evaluated in the program Tracer ver. 1.2.1, and as expected, many of these parameters did not reached stationarity even after five million sampling generations (data not shown). Under both models 7A and 7D the likelihood/sample generation plots reveal that convergence of likelihoods was not reached under either model after three million generations (Figs. 24-25). For model 7A, a plateau was reached between one and three million generations, but then a second plateau was reached between three and five million generations with the mean log likelihood lower than 50 (Fig. 24). Very similar results were recovered for model 7D (Fig. 25). This may suggest that too many model parameters are slowing the rate of convergence due to the large negative correlations among these parameters in the sampled posterior probability (Rannala, 2002). Nonetheless, it is unclear that the lack of convergence of model parameters, and the lack of a stabilized mean likelihood greatly affected the resulting trees, given their similarities to one another and their general agreement with parsimony.

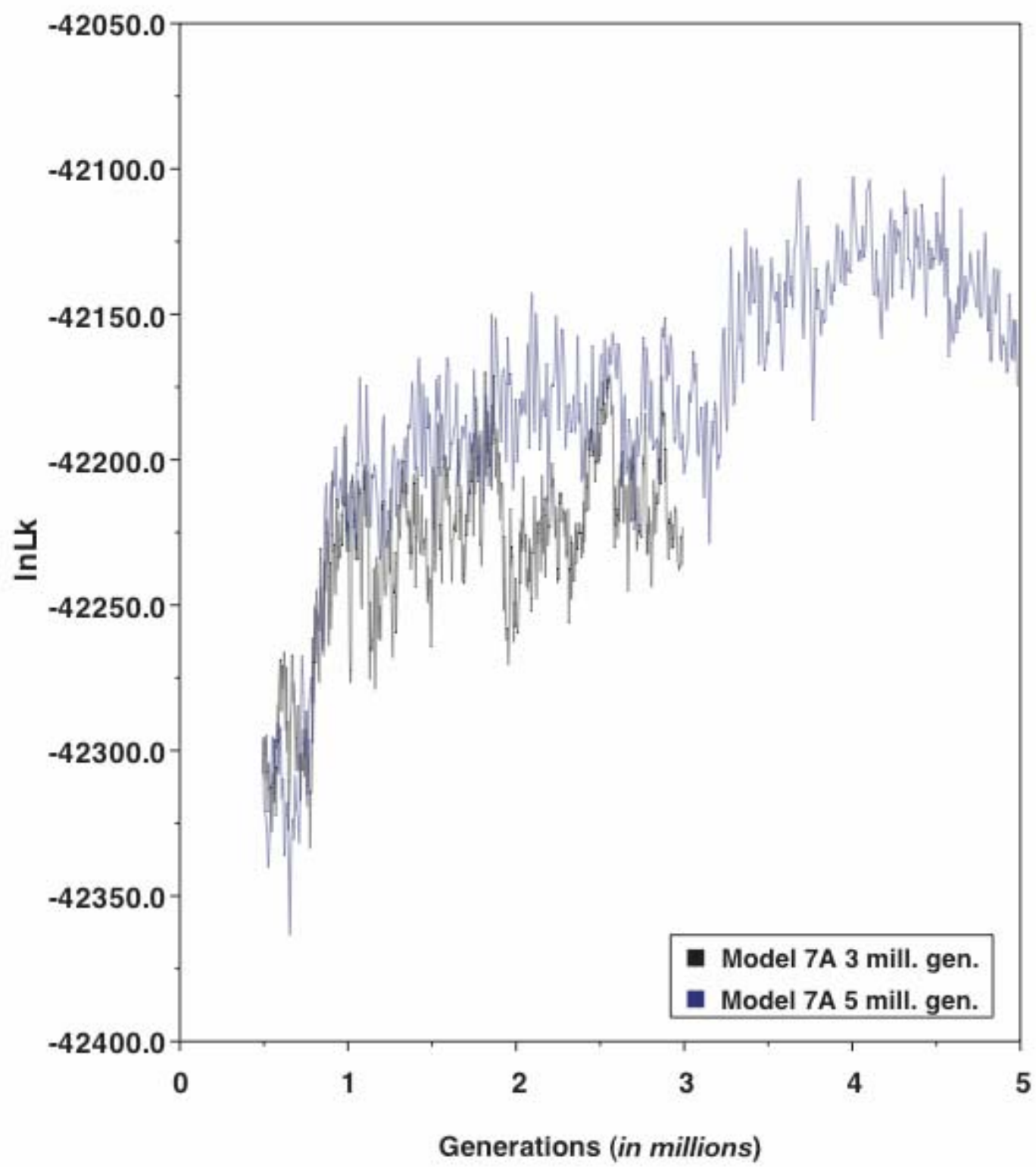


Figure 24. Plot of the log likelihood ($\ln L$) over the sampled generations for the three and five million generation analyses performed on the combined data with rRNA basepairs modeled under model 7A.

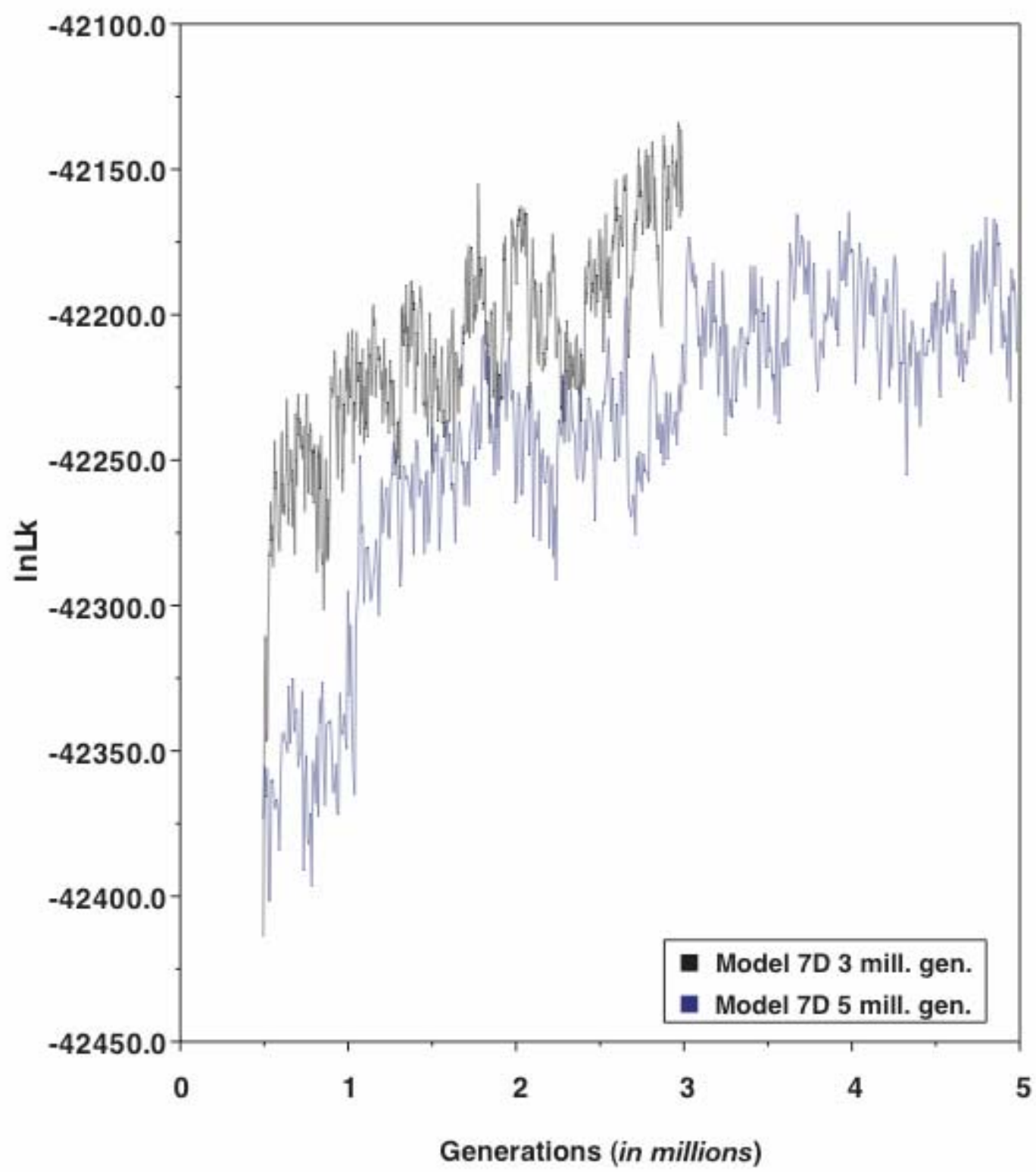


Figure 25. Plot of the log likelihood ($\ln L$) over the sampled generations for the three and five million generation analyses performed on the combined data with rRNA basepairs modeled under model 7D.

Galerucinae phylogeny

This phylogenetic analysis is the largest study on the historical relatedness of the galerucines and related leaf beetle kin, and is another step in the growing progress towards adequately sampling the entire major taxonomic delineations put forth by Seeno and Wilcox (1982). As with other studies (Reid, 1995a, b; Crowson & Crowson, 1996; Farrell, 1998; Lingafelter & Konstantinov, 2000; Gillespie *et al.*, 2003; Kim *et al.*, 2003; Duckett *et al.*, 2004; Gillespie *et al.*, 2004a), however, the gross under-sampling of certain lineages is likely complicating phylogeny estimation. This current study contributed many new sequences from Old World taxa and helped establish new monophyletic groups that were robust to the optimality criteria implemented here. For example, no previous analyses were able to confidently place the Oidini and Hylaspini within the Galerucinae *sensu stricto*. Also, although it was only recovered under parsimony, I establish for the first time the monophyly of all four diabroticine sections. I suspect that better efforts in future maximum likelihood analyses will eventually recover this result. Even with gross under-sampling of the luperine subtribe Luperina, several long-standing sections were recovered as monophyletic, namely Scelidites and Monoleptites. A third section, Phyllobroticites, was nearly monophyletic in all analyses (many instances of paraphyly) except for a few aberrant taxa, especially the two sampled species in the genus *Mimastra*. It is imperative that future studies include an increased taxon sampling of the subtribe Luperina in order to determine the relationships among the three subtribes, and the overall placement of the tribe within the other four tribes of the Galerucinae.

Several recent studies have addressed the evolution of cucurbitacin specialization in Old World Aulacophorina and New World Diabroticina within a phylogenetic framework (Gillespie 2001; Gillespie *et al.*, 2003, 2004a). I did not significantly improve the sampling of the genera *Aulacophorina* and *Acalymma*, the two genera with known cucurbit specialists. However, with the increased taxon sampling in both the Aulacophorina and Diabroticina, the placements of these genera are more solid. In no analysis under any optimality criteria are these two tribes remotely related. Furthermore, within the Diabroticina, *Acalymma* is never basal to the remaining taxa, adding strength to the argument that these beetles have independently gained the ability to sequester toxic cucurbitacins for selective benefits such as mating advantages and predator/pathogen avoidance (Tallamy *et al.* 1999; Gillespie *et al.*, 2003, 2004a). Only with adequate sampling of the Luperina, and hence a strongly supported phylogeny for all three subtribes of the Luperini, will this argument be entirely convincing.

Corroboration

As in previous studies on the phylogenetic reconstruction of this vast and intriguing beetle taxon, I have sought to combine approaches using different optimality criteria in an effort to let the data "speak for itself". The predictive power in phylogeny estimation is providing evidence that a result is robust to a variety of optimality criteria (e.g., Kjer *et al.*, 2001; Gillespie *et al.*, 2003, 2004a). In this study I demonstrate such corroboration by recovering many of the same monophyletic groups under parsimony and alternative likelihood effects can be attributed to gross under-sampling methods

Table 19. Monophyletic groups recovered in the parsimony and Bayesian (PHASE) analyses¹.

Taxon	TNT	RNA7A	RNA7D
Chrysomelinae + Trichostomata	yes	no	no
Chrysomelinae (- Timarchini) + Trichostomata	yes	yes	yes
Galerucinae <i>sensu lato</i>	yes	yes	yes
Galerucinae <i>sensu stricto</i>	yes	no	yes
Oidini	yes	yes	yes
Galerucini ²	yes	yes	yes
Galerucites	yes	yes	yes
Coelomerites	no	no	no
Alysites	no	no	no
Schematizites	no	no	no
Apophyllites	no	no	no
Metacyclini	no	no	no
New World genera	no	yes	yes
Old World genera	NA	NA	NA
Hylaspini ³	yes	yes	yes
Antiphites	no	no	no
Sermylites	no	no	no
Hylaspites	no	no	no
Agelastites	yes ⁴	yes ⁴	yes ⁴
Luperini	no	no	no
Aulacophorina	yes	yes	yes
Aulacophorites	yes	yes	yes
Idacanthites	NA	NA	NA
Diabroticina (all sections)	yes	no	no
Diabroticina (- <i>Trachyscelida</i>)	yes	yes	yes
Diabrotites ⁵	yes	yes	yes
<i>Acalymma sensu stricto</i>	yes	yes	yes
Cerotomites	yes ⁶	yes	yes
Phyllethrites	yes ⁷	yes ⁷	yes ⁷
Trachyscelidites	NA	NA	NA
Luperina	no	no	no
Adoxiites	NA	NA	NA
Scelidites	yes ⁸	yes ^{8,9}	yes ⁸
Phyllobrotites	no	no	no
Ornithognathites	NA	NA	NA
Exosomites	NA	NA	NA
Monoleptites	yes	yes	yes
Luperites	no	no	no

Table 19. Continued.

Taxon	TNT	RNA7A	RNA7D
Megalognathites	NA	NA	NA

¹ NA refers to single taxa that comprised a group in this study.

² Including Old World metacycline *Palaeophylia* sp.

³ Minus the two individuals of *Agelastica coerulea*.

⁴ Two individuals of *Agelastica coerulea* group together but not within the Hylaspini.

⁵ Minus *Gynandrobrotica* spp. (not including *G. nigrofasciata*).

⁶ Paraphyletic with Phyllethrites (in part).

⁷ Paraphyletic with Cerotomites (in part).

⁸ Inclusive of the exosomite *Pteleon brevicornis*.

⁹ Inclusive of the flea beetle *Orthaltica copalina*.

(Table 19). In most instances where the methods produced different results, the effects can be attributed to gross under-sampling of the questionable lineages. This likely implies that the sampled characters are providing a reasonable phylogenetic signal to elicit credible estimates of phylogeny. Thus inclusion of more taxa, especially in the under-sampled groups discussed above, will only improve the future analyses on this beetle group, and likely provide a phylogenetic hypothesis that will be credible enough to evaluate the entire taxonomic delineations proposed by Seeno and Wilcox (1982).

Experimental procedures

Taxa examined

Table 20 lists the chrysomeloid species analyzed in this investigation, with respective GenBank accession numbers for all sequences given. All newly collected data was added to existing character matrices from previous studies on this beetles group (Duckett

Table 20. A list of the 249 chrysomeloid taxa analyzed in this investigation.

Taxon ^a	Extract	Localtiy	Accession numbers		
(Family/Subfamily/Tribe/Subtribe/Section) Code ^b			COI	28S	18S
Orsodacnidae					
<i>Orsodacne atra</i> (Ahrens)	JJG114	United States: Utah: Tony Grove	AY242401	AY243660	
^k <i>Orsodacne atra</i> (Ahrens)	CND114	United States: Utah: Tony Grove	AY171396	AY171422	-----
Chrysomelidae					
Lamprosomatinae					
<i>Lamprosoma</i> sp. Kirby	JJG215	Brazil: Rio Grande do Sul: Porto Alegre	AY242392	AY243651	AY244878
Clytrinae					
<i>Clytrasoma palliatum</i>	JJG286	Unknown	-----	AY646286	AY244834
Criocerinae					
<i>Lema</i> sp. Fabricius	JJG308	Costa Rica: Guanacaste: Santa Elena	AY242400	AY243659	AY244835
<i>Lema</i> sp. Fabricius	JJG324	Costa Rica: Guanacaste: Santa Elena	-----	-----	
Cassidinae					
<i>Coptocycla adamantina</i> (Germar)	JJG214	Brazil: Rio Grande do Sul: Porto Alegre	AY242390	AY243649	AY244836
<i>Microhophala vittata</i> Baly	JJG218	United States: Utah: Logan Canyon	AY242391	AY243650	AY244837
Eumolpinae					
<i>Syneta</i> sp.	CND723	Korea: Kangwon	-----	AY646287	-----
^k <i>Syneta adamsi</i> Baly	SJK723	Korea: Kangwon	AY171414	AY171441	-----
<i>Megascelis</i> sp. Latreille	JJG244	Costa Rica: Heredia: ALAS	AY242393	AY243652	
<i>Metaxyonycha panamensis</i> Jacoby	JJG311			AY646288	AY244838
<i>Metaxyonycha</i> sp. Chevrolat	JJG132	Panama: Chiriqui: Continental Divide Trail, Fortuna	AY242394	AY243653	
<i>Callisina quadripustulata</i> Baly	JJG321	Thailand: Krabi	AY242395	AY243654	AY244839
<i>Colaspis</i> sp. Fabricius (or nr.)	JJG357	Peru: Loreto: Iquitos: Explor Tambos		AY646289	
<i>Colaspis</i> sp. Fabricius	JJG141	Panama: Chiriqui: Continental Divide Trail, Fortuna	AY242396	AY243655	
<i>Colasposoma</i> sp. Laporte	JJG318	Thailand: Khao Yai Nat'l. Pk.	AY242397	AY243656	AY244840
<i>Tymnes tricolor</i> (Fabricius)	JJG258	United States: North Carolina: Swain Co.	AY242398	AY243657	AY244841
<i>Chalcophana</i> sp. Chevrolat	JJG352	Costa Rica: Heredia: La Selva Biol. Sta.	AY242399	AY243658	AY244842
Chrysomelinae					
Chrysomelini					
<i>Chrysomela knabi</i> Brown	JJG237	United States: Utah: Benson	AY242402	AY243661	
<i>Chrysomela aeneicollis</i> (Schaeffer)	JJG277	Canada: Ontario: Alquonquin Provincial Pk.	AY242403	AY243662	
<i>Chrysomela populi</i> Linnaeus	JJG236	China: Beijing	AY242404	AY243663	
^k <i>Chrysomela tremulae</i> Fabricius	SJK705	France: Orleans	AY171397	AY171423	-----
^k <i>Chrysolina coerulans</i> (Scriba)	SJK703	France: Orleans	AY171403	AY171429	-----

Table 20. Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Locality	Accession numbers		
			COI	28S	18S
<i>Gastrophysa cyanea</i> Melsheimer	JJG329	United States: Texas: College Station	AY242405	AY243664	AY244843
^k <i>Paropsis porosa</i> Erichson	SJK704	Australia: Tasmania: Creekotow	AY171411	AY171438	-----
^k <i>Zygogramma piceicollis</i> (Stål)	CND334	United States: Arizona	AY171413	AY171440	-----
Timarchini					
<i>Timarcha</i> sp. Latreille	CND706	France: Vosges	-----	AY646290	AY244844
^k <i>Timarcha tenebricosa</i> (Fabricius)	SJK707	France: Vosges	AY171412	AY171439	-----
Galerucinae sensu lato					
Alticini					
^k <i>Altica</i> sp. Geoffroy	CND221	France: Montpellier	AY171398	AY171424	-----
^k <i>Allochroma</i> sp. Clark	CND327	Brazil: Areia Branca: PA	AY171428	AY171428	-----
^k <i>Aphthona nigriscutis</i> Foudras	SJK700	Russia: Krasnodar	AY171430	AY171430	-----
^k <i>Chaetocnema</i> sp. (Stephens)	SJK720	Korea: Chejudo	-----	AY171431	-----
(nr. <i>costulata</i>)					
^k <i>Disonycha conjuncta</i> (Germar)	CND061	Brazil: Rio Grande do Sul: Canguçu	AY171407	AY171434	-----
^k <i>Blepharida rhois</i> (Forster)	CND209	United States: New Jersey	AY171408	AY171435	-----
^k <i>Dibolia borealis</i> Chevrolat	CND419	United States: Tennessee	AY171415	AY171442	-----
^k <i>Sangariola fortunei</i> (Baly)	SJK721	Korea: Kyongbuk	AY171416	AY171443	-----
<i>Systema</i> sp. Chevrolat (nr. <i>lustrans</i>)	JJG219	Brazil: Rio Grande do Sul: Canguçu	AY242406	AY243665	AY244845
^k <i>Systema bifasciata</i> Jacoby	SJK219	Brazil: Rio Grande do Sul: Canguçu	AY171405	AY171432	-----
<i>Scelidopsis</i> sp. Jacoby	JJG225	Costa Rica: Guanacaste: Santa Elena	AY242407	AY243666	AY244854
<i>Cacoscelis</i> sp. Chevrolat	JJG195	Brazil: Paraná: Admirante Tamandaré	AY242408	AY243667	AY244855
<i>Epitrix fasciata</i> Blatchley	JJG328	United States: Texas: College Station	AY242409	AY243668	AY244848
<i>Physodactyla rubiginosa</i> (Gerstaecker)	CND253	South Africa: Kwa-Zulu Natal: Howick	AF479427	AY243671	AY244851
<i>Alagoasa libentina</i> (Germar)	CND303	Brazil: Sao Paulo: Sao Paulo	AF479470	AY243670	AY244850
<i>Walterianella bucki</i> Bechyné	CND039	Brazil: Parana: Piraquara, Manancias	AF479458	AY243673	AY244853
<i>Blepharida ornata</i> Baly	CND209	South Africa: Kwa-zulu Natal	AF479419	AY243672	AY244852
<i>Megistops vandepolli</i> Duvivier	CND002	Brazil: Rio Grande do Sul State	AY242410	AY243669	AY244849
<i>Luperaltica</i> sp. Crotch (or nr.)	JJG253	Costa Rica: Heredia: La Selva Biol. Sta.	AY242430	AY243695	AY244856
^k <i>Orithaltica copalina</i> (Fabricius)	SJK721	United States: North Carolina	AY171410	AY171437	-----
<i>Aedmon morrisoni</i> Blake	CND207	Puerto Rico: Salinas		AY646291	AY244857
<i>Crimissa curalis</i>	CND329	Brazil: Tocantins	-----	-----	AY244846
<i>Pseudadorium</i> sp.			-----	-----	
<i>Procalus mutans</i>	CND254	Chile: Isla Quiriquina	-----	-----	AY244847
Galerucinae sensu stricto					
Oidini					

Table 20. Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Locality	Accession numbers		
			COI	28S	18S
<i>Oides decempunctata</i> (Billberg)	JJG334	China: Beijing	AY242411	AY243674	AY244868
^k <i>Oides decempunctata</i> (Billberg)	SJK718	Korea: Chonbuk	AY171421	AY171448	-----
<i>Oides andrewsi</i> Jacoby	JJG409	Thailand: Kanchanaburi: Sri Sawat		AY646292	
<i>Oides andrewsi</i> Jacoby	JJG439	Thailand: Chiang Mai Prov: Chiang Dao	-----	AY646293	-----
<i>Oides andrewsi</i> Jacoby	JJG445	Thailand: Kanchanaburi: Throng Pha Phoom	-----	-----	
<i>Oides lividus</i>	JJG408	Thailand: Chiang Mai Prov		-----	-----
<i>Anoides</i> sp. Weise (or nr.)	JJG380	Papua New Guinea		AY646294	-----
Galerucini					
Galerucini Chapuis "genus undet."	JJG387	Papua New Guinea		AY646295	-----
Galerucites					
<i>Galeruca</i> sp. Geoffroy	CND700	United States: Utah	-----	AY646297	
^k <i>Galeruca rudis</i> LeConte	CND702	United States: Utah	AY171409	AY171436	-----
<i>Galeruca yubis</i>	JJG197	United States: Utah: Toney Grove	-----	-----	
Coelomerites					
<i>Caraguata pallida</i> (Jacoby) (or nr.)	JJG139	Panama: Panama: Gamboa	AY242510	AY243776	AY244875
<i>Dircema cyanipenne</i> Bechyné (or nr.)	JJG118	Peru: Loreto: Iquitos	AY242505	AY243771	AY244877
<i>Dircema</i> sp. Clark	JJG343	Panama: Panama: Cerro Campana	AY242506	AY243772	
<i>Dircema</i> sp. Clark (or nr.) marginatum gp.	JJG350	Costa Rica		AY646298	
<i>Dircema</i> sp. Clark evidens gp.	JJG355	Costa Rica		AY646299	
<i>Dircema</i> sp. Clark	JJG449	Costa Rica		AY646300	
<i>Dircemella</i> sp. Weise	JJG202	South Africa: Kwa-Zulu Natal, Bayala	AY242507	AY243773	
<i>Dircemella</i> sp. Weise	JJG307	South Africa: Kwa-Zulu Natal, Bayala	AY242508	AY243774	
<i>Trirhabda bacharidis</i> (Weber)	JJG075	United States: New Jersey: Ocean Co.	AY242503	AY243769	
^k <i>Monocesta</i> sp. Clark	CND710	United States	AY171406	AY171433	-----
<i>Cerochroa brachialis</i> Stål	JJG405	South Africa		AY646301	-----
<i>Sastroides</i> sp.	JJG438	Thailand: Chiang Mai Prov: Chiang Dao		-----	
Atysites					
<i>Diorhabda</i> sp. Weise	CND712	China: Beijing	AY242518	AY243784	-----
^k <i>Diorhabda elongata</i> (Brullé)	SJK712	United States	AY171419	AY171446	-----
<i>Megaleruca</i> sp. Laboisière	JJG204	South Africa: Kwa-Zulu Natal, Eshowe	AY242514	AY243780	
<i>Megaleruca</i> sp. Laboisière	JJG309	South Africa: Kwa-Zulu Natal, Bayala	AY242513	AY243779	
<i>Megaleruca</i> sp. Laboisière	JJG320	South Africa: Kwa-Zulu Natal, Eshowe		AY646302	
<i>Pyrrhalta maculicollis</i> (Motschulsky)	JJG190	China: Beijing	AY242515	AY243781	
<i>Pyrrhalta aenescens</i> (Fairmaire)	JJG187	France		AY646303	
<i>Pyrrhalta</i> sp. Joannis	JJG316	Thailand: Khao Yai Nat'l. Pk.	AY242516	AY243782	-----

Table 20. Continued.

Taxon ^a	Extract	Localtiy	Accession numbers		
(Family/Subfamily/Tribe/Subtribe/Section) Code ^b			COI	28S	18S
Schematizites					
<i>Metrogaleruca</i> sp. Bechyné & Bechyné	JJG134	Panama: Panama: Gamboa	AY242511	AY243777	
<i>Monoxia debilis</i> LeConte	JJG239	United States: Utah: Logan Co.	AY242512	AY243778	
<i>Neolachmaea dilatipennis</i> (Jacoby)	JJG323	Brazil: Rio Grande do Sul: Maquiné	AY242519	AY243785	
<i>Ophraea</i> sp. Jacoby (or. nr.)	JJG131	United States: undetermined SW locality	AY242504	AY243770	
<i>Ophraella notulata</i> (Fabricius)	JJG095	United States: New York: Suffolk Co.	AY242517	AY243783	
<i>Schematiza flavofasciata</i> (Klug)	JJG188	Brazil: Rio Grande do Sul: Canguçu	AY242520	AY243786	
^k <i>Schematiza flavofasciata</i> (Klug)	ZSH003	Brazil: Rio Grande do Sul: Canguçu	AY171420	AY171447	-----
<i>Erynephala punticollis</i>	JJG193	Unknown		-----	
Apophyllites (apo)					
<i>Pseudadimonia variolosa</i> (Hope)	JJG312	Thailand: Krabi	AY242509	AY243775	
<i>Apophyllia pallipes</i> (Baly)	JJG429	Thailand: Chiang Mai Prov: Doi Pui	-----	AY646304	
Metacyclini					
New World genera					
<i>Chthoneis</i> sp. Baly	JJG109	Costa Rica: Oratino: Carara	AY242498	AY243764	AY244872
<i>Chthoneis</i> sp. Baly (nr. <i>marginicollis</i>)	JJG354	Costa Rica		AY646305	
<i>Chthoneis</i> sp. Baly (nr. <i>iquitoensis</i>)	JJG361	Costa Rica		AY646306	
<i>Masurius violaceipennis</i> (Jacoby) (or nr.)	JJG116	Panama: Panama: Cerro Campana	AY242500	AY243766	AY244870
<i>Malachorhinus sericeus</i> Jacoby	JJG129	Costa Rica: Heredia: La Selva Biol. Sta.	AY242499	AY243765	AY244869
<i>Exora obsoleta</i> (Fabricius)	JJG110	Costa Rica: Guanacaste: San Luis Valley	AY242496	AY243762	
<i>Exora obsoleta</i> (Fabricius)	JJG353	Peru: Loreto: Iquitos	AY242497	AY243763	
<i>Exora</i> sp. Chevrolat	JJG340	Guatemala: Huchuctenango Prov: Barrillas		AY646307	
<i>Pyesia</i> sp. Clark	JJG246	Costa Rica: Heredia: ALAS	AY242501	AY243767	AY244873
<i>Zepherina</i> sp. Bechyné (or nr.)	JJG342	Peru: Loreto: Iquitos: Explororama lodge		AY646308	
Old World genus					
<i>Palaeophyllia</i> sp. Jacoby (or nr.)	JJG222	South Africa: Kwa-Zulu Natal, Eshowe	AY242502	AY243768	AY244874
Hylaspini					
Antiphites					
<i>Pseudeusthetia hirsuta</i>	JJG443	Thailand: Phu ket Prov: Ton Sai Waterfall		AY646309	
<i>Emathea subcaerulea</i>	JJG442	Thailand: Kanchanaburi: Sri Sawat		AY646310	
<i>Emathea subcaerulea</i>	JJG410	Thailand: Kanchanaburi: Sri Sawat		-----	-----
Sermylites					
<i>Aplosonyx orientalis</i> (Jacoby)	JJG436	Thailand: Chiang Mai Prov: Chiang Dao	-----	AY646311	
<i>Aplosonyx quadriplagiatus</i> (Baly)	JJG173	Sumatra: Aceh: Soraya Field Station	-----	AY243675	-----
<i>Aplosonyx</i> sp. Chevrolat	JJG427	Thailand: Chiang Mai Prov: Doi Pui	-----	AY646312	

Table 20. Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Locality	Accession numbers		
			COI	28S	18S
<i>Aplosonyx</i> sp. Chevrolat	JJG412	Thailand: Chiang Mai Prov		AY646313	-----
<i>Sermylassa halensis</i> (Linnaeus)	JJG179	France	-----	AY243676	-----
Hylaspites					
<i>Agelasa nigriceps</i> Motschulsky	JJG319	Japan: Honshu: Tamioka: Fukushima	AY242412	AY243677	
<i>Doryidella</i> sp. Laboissière (or nr.)	JJG425	Thailand: Kanchanaburi: Sri Sawat		AY646314	
<i>Sphenoraia paviei</i> Laboissière	JJG437	Thailand: Chiang Mai Prov: Chiang Dao	-----	AY646315	-----
Agelastites					
<i>Agelastica coerulea</i> Baly	JJG315	South Korea: Seoul	AY242413	AY243678	
^k <i>Agelastica coerulea</i> Baly	SJK701	South Korea: Seoul	AY171399	AY171425	-----
Luperini					
Luperini Chapuis "genus undet."	JJG376	Peru		AY646338	-----
Aulacophorina					
Aulacophorites					
<i>Paridea</i> sp. Baly (or nr.)	JJG235	Japan: Toyama, Kureya Hills	-----	AY243696	
<i>Chosnia obesa</i> (Jacoby) (or nr.)	JJG201	South Africa: Kwa-Zulu Natal, Eshowe	AY242431	AY243697	AY244862
<i>Sonchia sternalis</i> Fairmaire (or nr.)	JJG210	South Africa: Kwa-Zulu Natal, Bayala	AY242432	AY243698	AY244863
<i>Aulacophora indica</i> (Gmelin)	JJG220	Taiwan: Taipei	AY242435	AY243701	AY244864
^k <i>Aulacophora indica</i> (Gmelin)	SJK711	Taiwan: Taipei	AY171417	AY171444	-----
<i>Aulacophora lewisii</i> Baly	JJG158	Taiwan: Taipei	AY242434	AY243700	
<i>Aulacophora lewisii</i> Baly	JJG228	Taiwan: Taipei	AY242433	AY243699	
<i>Aulacophora lewisii</i> Baly	JJG127	Taiwan: Taipei		AY646316	
<i>Leptaulaca fissicollis</i> Thomson (or nr.)	JJG234	South Africa: Sirihni, Kruger Park	AY242437	AY243703	
<i>Diacantha fenestrata</i> Chapuis (or nr.)	JJG232	South Africa: Sirihni	AY242438	AY243704	
Idacanthites					
<i>Prosmidia conifera</i> Fairmaire (or nr.)	JJG212	South Africa: Sirihni, Kruger Park	AY242436	AY243702	
Diabroticina					
Diabroticites					
Diabroticites Chapuis "genus undet."	JJG345	Peru: Loreto: Iquitos: Explor Napo	-----	AY646339	
<i>Isotes marginatus</i>	JJG105	Costa Rica: Guanacaste: Santa Elena		-----	
<i>Isotes multipunctata</i> (Jacoby)	JJG300	Mexico: Veracruz: Orizaba	AY242457	AY243723	
<i>Isotes</i> sp. Weise	JJG145	Panama: Chiriqui: Continental Divide Trail, Fortuna	AY242458	AY243724	
<i>Isotes</i> sp. Weise	JJG349	Peru: Loreto: Iquitos	AY242456	AY243722	
<i>Isotes</i> sp. Weise	JJG351	Peru: Loreto: Iquitos	AY242454	AY243720	
<i>Isotes</i> sp. Weise	JJG363	Peru: Loreto: Iquitos	AY242455	AY243721	

Table 20. Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Locality	Accession numbers		
			COI	28S	18S
<i>Isotes</i> sp. Weise	JJG372	Panama: Darien Prov.	AY242459	AY243725	
<i>Isotes</i> sp. Weise	JJG373	Panama: Darien Prov.	AY242460	AY243726	-----
<i>Paranapiacaba tricineta</i> (Say)	JJG322	United States: New Mexico: Albuquerque	AY242487	AY243753	AY244867
<i>Paranapiacaba</i> sp. Bechyné	JJG094	Brazil: Rio Grande do Sul: San Francisco, de Paula	AY242486	AY243752	
<i>Acalymma vittatum</i> (Fabricius)	JJG413	United States: Delaware: Newcastle Co.		AY646317	-----
<i>Acalymma fairmairei</i> (Baly)	JJG016	Mexico: Veracruz: Xalapa	AY242442	AY243708	
<i>Acalymma bivittatum</i> (Fabricius)	JJG297	Cuba: Havana: La Vibora	AY242443	AY243709	
<i>Acalymma blomorum</i> Munroe & R. Smith (or nr.)	JJG229	Mexico: Veracruz: Xalapa	AY242444	AY243710	
<i>Acalymma trivittatum</i> (Mannerheim)	JJG059	Costa Rica: Heredia: La Selva Biol. Sta.	AY242445	AY243711	
<i>Acalymma hirtum</i> (Jacoby)	JJG053	Costa Rica: Puntarenas: San Isidro, Quizarra	AY242446	AY243712	
<i>Acalymma albidovittatum</i> (Baly)	JJG305	Brazil: Rio Grande do Sul: Passo Fundo	AY242447	AY243713	
<i>Acalymma</i> sp. Barber	JJG359	Peru: Loreto: Iquitos	AY242448	AY243714	
<i>Acalymma</i> sp. Barber	JJG360	Peru: Loreto: Iquitos	AY242449	AY243715	
<i>Acalymma</i> sp. Barber	JJG399	Puerto Rico: Carite Reserve		AY646318	
<i>Paratriarius subimpressa</i> (Jacoby)	JJG128	Costa Rica: Heredia: La Selva Biol. Sta.	AY242461	AY243727	
<i>Paratriarius</i> sp. Schaeffer	JJG147	Panama: Chiriqui: Continental Divide Trail, Fortuna	AY242462	AY243728	
<i>Paratriarius</i> sp. Schaeffer	JJG348	Peru: Loreto: Iquitos	AY242463	AY243729	
<i>Paratriarius</i> sp. Schaeffer	JJG374	Panama: Darien Prov.	AY242464	AY243730	
<i>Amphelasma nigrolineatum</i> (Jacoby)	JJG227	Costa Rica: Guanacaste: Santa Elena	AY242488	AY243754	
<i>Amphelasma sexlineatum</i> (Jacoby)	JJG295	Mexico: Tlaxcala: Cuernavaca	AY242489	AY243755	
<i>Diabrotica balteata</i> LeConte	JJG288	Mexico: Veracruz: Orizaba	AY242465	AY243731	
<i>Diabrotica biannularis</i> Harold	JJG010	Mexico: Veracruz: Xalapa	AY242466	AY243732	-----
<i>Diabrotica decempunctata</i> (Latreille)	JJG299	Panama: Darien Prov.	AY242467	AY243733	
<i>Diabrotica speciosa</i> (Germar)	JJG306	Unknown		AY646319	
<i>Diabrotica speciosa speciosa</i> (Germar)	JJG125	Brazil: Mato Grosso State: Sapazal Co.		AY271865	
<i>Diabrotica virgifera virgifera</i> LeConte	JJG060	United States: Delaware: Newcastle Co.	AY242468	AY243734	
<i>Diabrotica adelpha</i> Harold	JJG046	Costa Rica: Guanacaste: Santa Elena	AY242469	AY243735	
<i>Diabrotica porracea</i> Harold	JJG292	Mexico: Veracruz	AY242471	AY243737	
<i>Diabrotica undecimpunctata howardi</i> Barber	JJG370	United States: Texas: Bryan	AY242473	AY243739	-----
<i>Diabrotica undecimpunctata howardi</i> Barber	JJG223	United States: Delaware: Newcastle Co.	AY242472	AY243738	

Table 20. Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Localtiy	Accession numbers		
			COI	28S	18S
^k <i>Diabrotica undecimpunctata howardi</i> Barber	SJK223	United States: Delaware: Newcastle Co.		AY171445	-----
<i>Diabrotica tibialis</i> Jacoby	JJG170	Hondoras: Morazan: Zamorano	AY242480	AY243746	
<i>Diabrotica limitata</i> (Sahlberg)	JJG313	Brazil: Rio Grande do Sul: Serato	AY242481	AY243747	
<i>Diabrotica l. quindecimpunctata</i> (Germar)	JJG180	Brazil: Rio Grande do Sul	AY242470	AY243736	
<i>Diabrotica viridula</i> (Fabricius)	JJG314	Brazil: Rio Grande do Sul: Passo Fundo	AY242482	AY243748	-----
<i>Diabrotica rufolinbata</i>	JJG198	Brazil: Trinafu	-----	-----	
<i>Diabrotica graminea</i>	JJG398	Puerto Rico: Carite Reserve		-----	-----
<i>Diabrotica</i> sp. Chevrolat	JJG335	Guatemala: Guatemala City	AY242474	AY243740	
<i>Diabrotica</i> sp. Chevrolat	JJG336	Guatemala: Guatemala City	AY242475	AY243741	AY244865
<i>Diabrotica</i> sp. Chevrolat	JJG341	Guatemala: Huehuetenango	AY242476	AY243742	
<i>Diabrotica</i> sp. Chevrolat	JJG356	Peru: Loreto: Iquitos	AY242477	AY243743	
<i>Diabrotica</i> sp. Chevrolat	JJG362	Peru: Loreto: Iquitos	AY242478	AY243744	
<i>Diabrotica</i> sp. Chevrolat	JJG365	Peru: Loreto: Iquitos	AY242479	AY243745	-----
<i>Gynandrobrotica nigrofasciata</i> (Jacoby)	JJG152	Costa Rica: Guanacaste: San Luis Valley	AY242451	AY243717	
<i>Gynandrobrotica lepida</i> (Say)	JJG298	Costa Rica: Guanacaste: San Luis Valley	AY242452	AY243718	
<i>Gynandrobrotica</i> sp. Bechyné	JJG358	Peru: Loreto: Iquitos	AY242450	AY243716	
<i>Gynandrobrotica</i> sp. Bechyné	JJG371	Panama: Panama: Cerro Campana	AY242453	AY243719	
<i>Gynandrobrotica ventricosa</i> (Jacoby)	JJG135	Costa Rica		AY646321	
Cerotomites					
<i>Neobrotica caeruleofasciata</i> Jacoby	JJG117	Costa Rica: Puntarenas: San Isidro del el General	AY242483	AY243749	
<i>Neobrotica</i> sp. Jacoby	JJG337	Guatemala: Guatemala City	AY242484	AY243750	
<i>Neobrotica</i> sp. Jacoby	JJG375	Panama: Darien Prov.	AY242485	AY243751	
<i>Eucerotoma</i> sp. Laboissière	JJG344	Peru: Loreto: Iquitos	AY242490	AY243756	
<i>Eucerotoma</i> sp. Laboissière	JJG346	Peru: Loreto: Iquitos	AY242493	AY243759	
<i>Eucerotoma</i> sp. Laboissière	JJG347	Peru: Loreto: Iquitos	AY242491	AY243757	
<i>Eucerotoma</i> sp. Laboissière	JJG364	Peru: Loreto: Iquitos	AY242492	AY243758	
<i>Cerotoma arcuata</i> (Olivier)	JJG048	Brazil: Paraná: Londrina	AY242494	AY243760	
<i>Cerotoma</i> sp. Chevrolat	JJG339	Guatemala: Guatemala City	AY242495	AY243761	
<i>Cerotoma ruficornis</i> (Olivier)	JJG172	Costa Rica	-----	AY646322	
<i>Cerotoma ruficornis</i> (Olivier)	JJG108	Costa Rica: Guanacaste: Santa Elena	-----	-----	-----
<i>Cerotoma facialis</i> Erichson	JJG161		-----	AY646323	
Phyllecthrites					
<i>Trichobrotica nymphaea</i> Jacoby	JJG226	Costa Rica: Guanacaste: Santa Elena	AY242440	AY243706	

Table 20. Continued.

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Locality	Accession numbers		
			COI	28S	18S
<i>Phyllethriss gentilis</i> LeConte	JJG366	United States: Missouri: Camden Co.	AY242441	AY243707	
<i>Phyllethriss</i> Dejean "genus undet."	JJG377	Peru		AY646324	
Trachyscelidites					
<i>Trachyscelida</i> sp. Horn	JJG224	Costa Rica: Heredia: La Selva Biol. Sta.	AY242439	AY243705	
Luperina					
Adoxiites					
<i>Medythia suturalis</i> (Motschulsky)	JJG434	Thailand: Kanchanaburi: Throng Pha Phoom	-----	AY646325	
<i>Medythia suturalis</i> (Motschulsky)	JJG448	Thailand: Chiang Mai Prov: Chiang Mai	-----	AY646326	-----
Scelidites					
<i>Scelolyperus lecontii</i> (Crotch)	JJG099	United States: Wyoming: Grand Teton Nat'l. Pk.	AY242419	AY243684	
<i>Scelolyperus meracus</i> (Say)	JJG257	United States: North Carolina: Swain Co.	AY242421	AY243686	
<i>Scelolyperus</i> sp. Crotch	JJG054	United States: Montana	AY242420	AY243685	
<i>Lygistus streptophallus</i> Wilcox	JJG367	United States: Arizona	AY242422	AY243687	
<i>Keithaeus blakeae</i> (White)	JJG414	United States: Texas: Chisos Lodge		AY646327	
<i>Stenoluperus nipponensis</i> Laboissière	CND717	South Korea: Chejudo Prov.	AY242429	AY243694	-----
Phyllobroticites					
<i>Phyllobrotica</i> sp. Chevrolat	JJG076	United States: Delaware: Newcastle Co.	AY242425	AY243690	
^b <i>Phyllobrotica</i> sp. Chevrolat	SJK076	United States: Delaware: Newcastle Co.		AY171427	-----
<i>Mimastra gracilicornis</i> Jacoby	JJG287	Thailand: Chiang Mai Prov: Doi Inthanon Nat'l. Pk.	AY242426	AY243691	
<i>Mimastra</i> sp. Baly	JJG430	Thailand: Chiang Mai Prov: Doi Pui	-----	AY646328	
<i>Hoplasoma unicolor</i> Illiger	JJG419	Thailand: Chiang Mai Prov: Chiang Dao		AY646329	
Ornithognathites					
<i>Hallirhotius</i> sp. Jacoby	JJG206	South Africa: Kwa-Zulu Natal, Eshowe	AY242424V	AY243689	AY244860
Exosomites					
<i>Pteleon brevicornis</i> (Jacoby)	JJG415	United States: Texas: Window Trail		AY646330	
<i>Liroetiella bicolor</i> Kimoto	JJG368	Thailand: Chiang Mai Prov: Doi Inthanon Nat'l. Pk.		AY646331	
<i>Cassena indica</i> (Jacoby)	JJG416	Thailand: Chiang Mai Prov: Doi Inthanon Nat'l. Pk.		AY646332	
Monoleptites					
<i>Desbordesius</i> sp.	JJG406	Thailand: Krabi		-----	
Monoleptites Chapuis "genus undet."	JJG422	Thailand: Chiang Mai Prov: Chiang Dao		AY646333	
Monoleptites Chapuis "genus undet."	JJG431	Thailand: Chiang Mai Prov: Doi Pui	-----	AY646334	

Table 20. Continued

Taxon ^a (Family/Subfamily/Tribe/Subtribe/Section) Code ^b	Extract	Locality	Accession numbers		
			COI	28S	18S
Monoleptites Chapuis "genus undet."	JJG440	Thailand: Kanchanaburi: Throng Pha Phoom		AY646335	-----
Monoleptites Chapuis "genus undet."	JJG338	Guatemala: 10K N of Solome		AY646296	
<i>Monolepta nigrotibialis</i> Jacoby	JJG044	South Africa: Umtumvuna	AY242416	AY243681	
^k <i>Monolepta nigrotibialis</i> Jacoby	SJK044	South Africa: Umtumvuna	AY171400	AY171426	-----
<i>Monolepta</i> sp. Chevrolat	JJG183	Japan: Toyama: Ida River	AY242417	AY243682	AY244861
<i>Monolepta</i> sp. Chevrolat	JJG310	South Africa: Kwa-Zulu Natal, St. Lucia	AY242414	AY243679	
<i>Monolepta</i> sp. Chevrolat	JJG317	South Africa: Sirihni, Kruger Park	AY242415	AY243680	
<i>Monolepta</i> sp. Chevrolat	JJG369	South Africa: Sirihni, Kruger Park	AY242418	AY243683	
<i>Metrioidea</i> sp. Fairmaire (or nr.)	JJG301	Brazil: Santa Catartina: Morro Bau	AY242423	AY243688	
Luperites					
<i>Spilocephalus bipunctatus</i> Allard	JJG205	South Africa: Kwa-Zulu Natal, Eshowe	AY242427	AY243692	
<i>Palpoxena</i> sp. Baly	JJG230	South Africa: Kwa-Zulu Natal, St. Lucia	AY242428	AY243693	
<i>Luperus longicornis</i> Fabricius	JJG407	Scotland: Gatehouse of Fleet	-----	AY646336	
Megalognathites					
<i>Megalognatha</i> sp. Baly	JJG303	South Africa	-----	AY646337	-----
<i>Megalognatha festiva</i> pale variety	JJG400	South Africa		-----	-----
<i>Megalognatha festiva</i> dark variety	JJG401	South Africa		-----	
Unidentified specimens					
Thailand specimen 4	JJG411	Thailand: Kanchanaburi: Sri Sawat	-----	AY646340	
Thailand specimen 7	JJG417	Thailand: Chiang Mai Prov: Chiang Dao		AY646341	
Thailand specimen 8	JJG418	Thailand: Chiang Mai Prov: Chiang Dao		AY646342	
Thailand specimen 10	JJG420	Thailand: Chiang Mai Prov: Chiang Dao		AY646343	
Thailand specimen 11	JJG421	Thailand: Chiang Mai Prov: Doi Inthanon Nat'l. Pk.		AY646344	
Thailand specimen 13	JJG423	Thailand: Kanchanaburi: Sri Sawat		AY646345	
Thailand specimen 14	JJG424	Thailand: Kanchanaburi: Sri Sawat	-----	AY646346	
Thailand specimen 22	JJG432	Thailand: Chiang Mai Prov: Doi Inthanon Nat'l. Pk.	-----	AY646347	
Thailand specimen 25	JJG435	Thailand: Kanchanaburi: Throng Pha Phoom	-----	AY646348	
Thailand specimen 31	JJG441	Thailand: Chiang Mai Prov: Chiang Dao		AY646349	
Thailand specimen 34	JJG444	Thailand: Chiang Mai Prov: Chiang Dao		-----	
Thailand specimen 36	JJG446	Thailand: Kanchanaburi: Throng Pha Phoom	-----	AY646350	
Thailand specimen 37	JJG447	Thailand: Chiang Mai Prov: Chiang Mai:		AY646351	

^a Taxonomic groupings follow Seeno & Wilcox (1982).
^b DNA extraction codes for all taxa are listed as recorded on all vouchered specimens.
^k Denotes sequence from Kim *et al.* (2003).

et al. 2004; Gillespie *et al.*, 2003, 2004a, b). Voucher specimens for all sampled taxa can be found in the Texas A&M University, Rutgers University, or the University of Delaware insect museums.

Genome isolation, PCR, and sequencing

For the sequences generated in this study, total genomic DNA was isolated using DNeasy™ Tissue Kits (Qiagen). PCR conditions followed those of Cognato & Vogler (2001), with primers designed for amplification of both the 28S-D2 and COI gene regions found in Gillespie *et al.* (2003, 2004a). Primers used for the 28S-D3 and 18S V4, V7-V9 are from Whiting *et al.*, 1997. I also designed internal primers to amplify and sequence taxa for which previously designed primers failed. All primers used in this study are listed in Table 21. Double-stranded DNA amplification products were sequenced directly with ABI PRISM™ (Perkin-Elmer) Big Dye Terminator Cycle Sequencing Kits and analyzed on an Applied Biosystems (Perkin-Elmer) 377 automated DNA sequencer. Both anti-sense and sense strands were sequenced for all taxa, and edited manually with the aid of Sequence Navigator™ (Applied Biosystems). During editing of each strand, nucleotides that were readable, but showed either irregular spacing between peaks, or had some significant competing background peak, were coded with lower case letters or IUPAC-IUB ambiguity codes. Consensus sequences were exported into Microsoft Word™ (Redmond, WA) or MacClade (version 4.0, Maddison & Maddison, 2000) for manual alignment.

Table 21. A list of the oligonucleotide primers used to amplify and sequence the chrysomeloids analyzed in this investigation¹.

CO1 primers:

CO1-1709Fs	sense	5'-TAA TTG GAG GAT TTG GAA ATT G-3'
C1-J-1751F	sense	5'-GGA TCA CCT GAT ATA GCA TTC CC-3'
CO1-1856F	sense	5'-ACN GGN TGA ACT GTY TAY CC-3'
C1-N-2191R	antisense	5'-CCC GGT AAA ATT AAA ATA TAA ACT TC-3'
CO1-2209R	antisense	5'-GAG AAA TTA TTC CAA ATC CRG GTA A-3'
CO1-2278R	antisense	5'-GCT AAT ATN GCA TAA ATT ATY CCY AA-3'

28S rRNA primers:

D2 UP-4	sense	5'-GAG TTC AAG AGT ACG TGA AAC CG-3'
D2UP COL1	sense	5'-CCG TTG AGG GGT AAA CCT GAG AAA C-3'
D2UP COL2	sense	5'-GGT AAA CCT GAG AAA CCC GAA A-3'
28S forward	sense	5'-GAG AGT TMA ASA GTA CGT GAA AC-3'
D2 DN-B	antisense	5'-CCT TGG TCC GTG TTT CAA GAC-3'
28SA	antisense	5'-CCT GAC TTC GTC CTG ACC AGG C-3'
28S-B (DN)	antisense	5'-TCG GAR GGA ACC AGC TAC TA-3'

18S rRNA primers:

(V4)

18Sa 0.7	sense	5'- ATT AAA GTT GTT GCG GTT-3'
18S CREM	sense	5'- CTT GAT TCG GTG TGG TGG TGC-3'
18Sb 2.5	antisense	5'-TCT TTG GCA AAT GCT TTC GC-3'
18S WALL1	antisense	5'-TTC AGT GTA GCG CGC GTG CGG CCC-3'
18S WALL2	antisense	5'-ATC ACA GAC CTG TTA TTG CTC-3'

(Domain III)

18Sa 2.0	sense	5'-ATG GTT GCA AAG CTG AAA C-3'
18S 3.I	antisense	5'-CAC CTA CGG AAA CCT TGT TAC GAC-3'

¹ The same primers were used for PCR and cycle-sequencing.

Multiple sequence alignment

I used MacClade v. 4.0 to color-code the COI sequences by translated amino acid state (using the *Drosophila* mitochondria code) to check for stop codons and gap-induced shifts to the reading frame. In an effort to determine whether or not to include

information from translated amino acid states, each amino acid substitution was evaluated within the predicted global insect structural model for COI (Lunt *et al.*, 1996). The rRNA sequences were aligned manually according to secondary structure, with the notation following Kjer *et al.* (1994) and Kjer (1995), with slight modifications (Gillespie *et al.*, 2004b). The D2-D3 alignment from Gillespie *et al.* (2004) was unchanged. Alignment of the 18S rRNA initially followed the secondary structural models of *Drosophila melanogaster* (Cannone *et al.*, 2002), with refinement to the variable region 4 (Van de Peer *et al.*, 1999) made from the double pseudoknot model of Wuyts *et al.* (2000). All regions variable in sequence length and base composition, especially hairpin-stem loops, were evaluated in the program *mfold* (version 3.1; <http://bioinfo.math.rpi.edu>), which folds RNA based on free energy minimizations (Mathews *et al.*, 1999; Zuker *et al.*, 1999). These free energy-based predictions were used to facilitate the search for potential base-pairing helices, which were confirmed only by the presence of compensatory base changes across a majority of taxa. Regions in which positional homology assignments were ambiguous across all taxa were defined according to structural criteria as in Kjer (1997) and characterized as regions of alignment ambiguity (RAA), regions of slipped-strand compensation (RSC) or regions of expansion and contraction (REC), following the methodology of Gillespie (2004). All of these unaligned regions were enclosed within brackets. Finally, a pairing mask (see Jow *et al.*, 2002; Hudelot *et al.*, 2003) was added to the alignment identifying basepairs within helices and unpaired regions.

Scripted manipulation

The three data partitions were combined into a NEXUS file for execution in PAUP* (version 4.0b10, Swofford, 1999). This file was used for all subsequent manipulations of the data. In addition, a helix index was created that identifies all pairing regions identified by the paring mask, as well as non-paired sites and bracketed regions. The Jrna scripts (available at the jRNA website: <http://hymenoptera.tamu.edu>) were used to integrate information from the NEXUS file and the helix index to create the following files: 1. individual NEXUS files for each data partition, 2. input file for the program TNT (Goloboff, 1997; Goloboff *et al.*, 2003), 3. input and control files for the program PHASE ver. 1.1 (Jow *et al.*, 2002; Hudelot *et al.*, 2003), 4. command line and input files delimited by secondary structure for the program POY (Gladstein & Wheeler, 1997), and 5. input file for the program MrBayes ver. 3.1 (Ronquist & Huelsenbeck, 2003). The jRNA scripts were also used to generate HTML formatted, color-highlighted alignments, summary statistics on basepair composition and covariation, basepair frequency tables, and column and region base composition. All of the abovementioned files and statistics are available at the jRNA website.

Model selection

Using the jRNA scripts I created four Nexus files separating paired and unpaired regions of the rRNA partitions. The files containing the unpaired regions of the alignment were analyzed in ModelTest (Posada & Crandall, 1998) to provide the best model of evolution for these positions in the rRNAs. For the 28S rRNA, the model of Tamura and Nei

(TN93, 1993) was reported as the best model under the hierarchical likelihood ratio test (hLRT, Posada & Crandall, 1998, 2001) and second best to model HKY85 under the Akaike information criterion (AIC, Linhart & Zucchini, 1986). Both the hLRT and AIC reported model TN93 as the best model for the 18S rRNA, and given that TN93 distinguishes between different transition classes, I elected to choose it over the simpler model HKY85. Thus all models analyzed in PHASE and MrBayes implemented the TN93+gamma+invariant sites model for non-pairing regions of the rRNA alignments. By partitioning the COI nucleotides into first, second and third codon positions, I exported three individual files per position from PAUP to evaluate the best model of evolution for each codon site. In ModelTest the hLRT and AIC reported different best models for the first and second codon positions: TrN+I+G (hLRT) and GTR+I+G (AIC) for the first position, and TVM+I+G (hLRT) GTR+I+G (AIC) for the second position. Both test statistics reported the most general time reversible model without a proportion of invariant sites (GTR+G) for the third positions, as all 155 third position sites contain at least some degree of base substitution. For the third and second positions I elected to use the most general time reversible model with invariant sites to err on the side of overparamterization rather than fitting the hypervariable positions to a simpler model of substitution. The results of Modeltest from all sub-partitions are listed in Table 22.

Table 22. Results from running Modeltest on the molecular partitions.

Partition	hLRTs	AIC
COI (1st pos.)	Model selected: TrN+I+G	Model selected: GTR+I+G
	-lnL = 5075.2705	-lnL = 5057.2578
	K = 7	K = 10
		AIC = 10134.5156
	Base frequencies:	Base frequencies:
	freqA = 0.3260	freqA = 0.3301
	freqC = 0.1754	freqC = 0.1907
	freqG = 0.1566	freqG = 0.1497
	freqU = 0.3421	freqU = 0.3295
	Substitution model:	Substitution model:
	Rate matrix	Rate matrix
	R(a) [A-C] = 1.0000	R(a) [A-C] = 0.2664
	R(b) [A-G] = 3.5142	R(b) [A-G] = 2.0363
	R(c) [A-U] = 1.0000	R(c) [A-U] = 0.6631
	R(d) [C-G] = 1.0000	R(d) [C-G] = 0.0000
	R(e) [C-U] = 25.2032	R(e) [C-U] = 12.8419
	R(f) [G-U] = 1.0000	R(f) [G-U] = 1.0000
	Among-site rate variation	Among-site rate variation
	Proportion of invariable sites (I) = 0.4257	Proportion of invariable sites (I) = 0.4322
	Variable sites (G)	Variable sites (G)
	Gamma distribution shape parameter = 0.4732	Gamma distribution shape parameter = 0.4925

Table 22. Continued.

Partition	hLRTs	AIC
COI (2nd pos.)	Model selected: TVM+I+G	Model selected: GTR+I+G
	-lnL = 1130.2550	-lnL = 1127.5142
	K = 9	K = 10
		AIC = 2275.0283
	Base frequencies:	Base frequencies:
	freqA = 0.2176	freqA = 0.1959
	freqC = 0.2137	freqC = 0.2678
	freqG = 0.1804	freqG = 0.1384
	freqU = 0.3883	freqU = 0.3978
	Substitution model:	Substitution model:
	Rate matrix	Rate matrix
	R(a) [A-C] = 6.4696	R(a) [A-C] = 2.2916
	R(b) [A-G] = 11.9582	R(b) [A-G] = 14.1500
	R(c) [A-U] = 2.3666	R(c) [A-U] = 1.4388
	R(d) [C-G] = 14.1298	R(d) [C-G] = 7.9184
	R(e) [C-U] = 11.9582	R(e) [C-U] = 4.0171
	R(f) [G-U] = 1.0000	R(f) [G-U] = 1.0000
	Among-site rate variation	Among-site rate variation
	Proportion of invariable sites (I) = 0.5965	Proportion of invariable sites (I) = 0.5069
	Variable sites (G)	Variable sites (G)
	Gamma distribution shape parameter = 0.4121	Gamma distribution shape parameter = 0.2665

Table 22. Continued.

Partition	hLRTs	AIC
COI (3rd pos.)	Model selected: GTR+G -lnL = 20507.4707 K = 9 Base frequencies: freqA = 0.1992 freqC = 0.1579 freqG = 0.0212 freqU = 0.6216 Substitution model: Rate matrix R(a) [A-C] = 0.2155 R(b) [A-G] = 9.2676 R(c) [A-U] = 0.1991 R(d) [C-G] = 1.1200 R(e) [C-U] = 1.6317 R(f) [G-U] = 1.0000 Among-site rate variation Proportion of invariable sites = 0 Variable sites (G) Gamma distribution shape parameter = 1.0452	Model selected: GTR+G -lnL = 20507.4707 K = 9 AIC = 41032.9414 Base frequencies: freqA = 0.1992 freqC = 0.1579 freqG = 0.0212 freqU = 0.6216 Substitution model: Rate matrix R(a) [A-C] = 0.2155 R(b) [A-G] = 9.2676 R(c) [A-U] = 0.1991 R(d) [C-G] = 1.1200 R(e) [C-U] = 1.6317 R(f) [G-U] = 1.0000 Among-site rate variation Proportion of invariable sites = 0 Variable sites (G) Gamma distribution shape parameter = 1.0452

Table 22. Continued.

Partition	hLRTs	AIC
18S (loops)	Model selected: TrN+I+G -lnL = 1643.3927 K = 7	Model selected: TrN+I+G -lnL = 1643.3927 K = 7 AIC = 3300.7854
	Base frequencies: freqA = 0.3068 freqC = 0.2193 freqG = 0.2097 freqU = 0.2642	Base frequencies: freqA = 0.3068 freqC = 0.2193 freqG = 0.2097 freqU = 0.2642
	Substitution model: Rate matrix R(a) [A-C] = 1.0000 R(b) [A-G] = 1.1110 R(c) [A-U] = 1.0000 R(d) [C-G] = 1.0000 R(e) [C-U] = 6.7264 R(f) [G-U] = 1.0000	Substitution model: Rate matrix R(a) [A-C] = 1.0000 R(b) [A-G] = 1.1110 R(c) [A-U] = 1.0000 R(d) [C-G] = 1.0000 R(e) [C-U] = 6.7264 R(f) [G-U] = 1.0000
	Among-site rate variation Proportion of invariable sites (I) = 0.6301 Variable sites (G) Gamma distribution shape parameter = 0.6021	Among-site rate variation Proportion of invariable sites (I) = 0.6301 Variable sites (G) Gamma distribution shape parameter = 0.6021

Table 22. Continued.

Partition	hLRTs	AIC
28S (loops)	Model selected: TrN+I+G -lnL = 2549.0569 K = 7 Base frequencies: freqA = 0.4050 freqC = 0.1879 freqG = 0.1643 freqU = 0.2427 Substitution model: Rate matrix R(a) [A-C] = 1.0000 R(b) [A-G] = 2.5251 R(c) [A-U] = 1.0000 R(d) [C-G] = 1.0000 R(e) [C-U] = 3.8543 R(f) [G-U] = 1.0000 Among-site rate variation Proportion of invariable sites (I) = 0.4302 Variable sites (G) Gamma distribution shape parameter = 0.5798	Model selected: HKY+I+G -lnL = 2549.9614 K = 6 AIC = 5111.9229 Base frequencies: freqA = 0.3893 freqC = 0.2052 freqG = 0.1521 freqU = 0.2534 Substitution model: Ti/tv ratio = 1.4141 Among-site rate variation Proportion of invariable sites (I) = 0.4569 Variable sites (G) Gamma distribution shape parameter = 0.6460

Because I could not determine a better model of evolution for the basepairs of the 28S rRNA in Chapter IV, I chose to combine the above-modeled partitions with both models 7A and 7D in a combined five-model analysis of the data. I concluded that I would determine the "better" analysis by analyzing the plots of likelihoods, tree lengths, and all model parameters in the program Tracer ver. 1.2.1 (Rambaut & Drummond, 2005) upon termination of the analyses, as well as through the corroboration of clades with those recovered by parsimony.

Phylogeny estimation

Parsimony analyses were performed with the programs PAUP* ver. 4.0b10 (Swofford, 2001) and TNT (Goloboff, 1999; Goloboff *et al.*, 2003). Search strategies followed those of Gillespie *et al.* (2004a). Nodal support was measured using the bootstrap (Felsenstein, 1985), performing 100 replicates with a cut-off of 50 percent.

Bayesian analysis under maximum likelihood was performed in two different programs; MrBayes, and PHASE. In MrBayes, the COI nucleotides were divided into their respective codon positions, as were the rRNA stems and loops. The covarion model was applied for the rRNA basepairs. Three independent analyses were run, each starting at a different seed. I used flat priors for all analyses. Six Markov chains, each run at different temperatures, were used in an effort to decrease time till convergence (Ronquist & Huelsenbeck 2003). I sampled every 1000th generation over sampling iterations of 3000, 6000 and 10000 generations.

In PHASE, the four best models reported above from Modeltest were combined with models 7A and 7D in two separate analyses. These analyses were run for three million and five million generations. To insure adequate mixing between the Markov chains, the three and five million generations were performed with different starting seeds. Like MrBayes, PHASE analyzes the data under maximum likelihood using Bayesian inference. I sampled every 1000 generations throughout each analysis using six Markov chains, keeping all chains at the same temperature and saving all branch lengths throughout. I used flat priors for all analyses. All analyses were performed on Xblast (Texas A&M University), a 21 compute element (42 cpus) cluster of Apple G4 Xserves running iNquiry (<http://xblast.tamu.edu/>). Initial analyses were performed to determine the burn-in, or time until an acceptable plateau is reached in the sampling of likelihoods, trees and parameters in the posterior probability distribution. These burn-in values were determined by plotting log likelihoods ($-\ln L$) and tree lengths (TL) over generation number in the program Tracer ver. 1.2.1 (Rambaut & Drummond, 2005). Ultimately, a highly conservative value of 500,000 generations was selected for the burn-in prior to each of the twelve analyses.

Parsimony analysis was also done using the program POY (Gladstein & Wheeler 1997). POY performs the processes of alignment and tree reconstruction simultaneously, in a procedure that leaves no statement of homologies other than the final tree. Although relatively recent in its development, POY actually approximates the Sankoff algorithm for simultaneous alignment and tree generation under parsimony (Sankoff *et al.*, 1973; Sankoff, 1975; Sankoff & Cedergren, 1983), something its writers

seemingly always forget to cite. Here POY analyses involve all combinations of bracketed and unbracketed data (as delimited by secondary structure), with both fixed states optimization (Wheeler, 1999) and direct optimization (Wheeler, 1996) tested under ts/tv/gap cost ratios of 1:1:1 and 2:1:1. Search strategies used in the POY analyses are the same as those performed in Gillespie *et al.* (2005).

Model evaluation

Results files from all Bayesian analyses were modified with the Jrna scripts to produce input files for Tracer. To determine that both analyses per model reached a similar sampling space in the posterior distribution, analyses of all sampling iterations were compared in Tracer. The recovery of similar results for tree lengths and topologies, clade posterior probabilities and parameter posterior probabilities from these iterations is a good indicator that stationarity has been reached and that the Markov sampling procedure is effectively sampling these statistics throughout the estimated sample sizes (ESS) (Huelsenbeck *et al.* 2002; Miller *et al.* 2004).

CHAPTER VI

ASSESSING THE ODD STRUCTURAL PROPERTIES OF NUCLEAR SMALL SUBUNIT RIBOSOMAL RNA SEQUENCES (18S) OF THE TWISTED-WING PARASITES (INSECTA: STREPSIPTERA)*

Overview

We report the entire sequence (2864 nts) and secondary structure of the nuclear small subunit ribosomal RNA (SSU rRNA) gene (18S) from the twisted-wing parasite *Caenocholax fenyasi texensis* Kathirithamby & Johnston (Strepsiptera: Myrmecolacidae). The majority of the base pairings in this structural model map onto the SSU rRNA secondary and tertiary helices that were previously predicted with comparative analysis. These regions of the core rRNA were unambiguously aligned across all Arthropoda. In contrast, many of the variable regions, as previously characterized in other insect taxa, had very large insertions in *C. f. texensis*. The helical basepairs in these regions were predicted with a comparative analysis of a multiple sequence alignment (that contains *C. f. texensis* and 174 published arthropod 18S rRNA

* This article, Gillespie, J.J., McKenna, C.H., Yoder, M.J., Gutell, R.R., Johnston, J.S., Kathirithamby, J., Cognato, A.I. Assessing the odd secondary structural properties of nuclear small subunit ribosomal RNA sequences (18S) of the twisted-wing parasites (Insecta: Strepsiptera). *Insect Mol Biol*, In press, is reprinted with permission from Blackwell Publishing, copyright 2005.

sequences, including 11 strepsipterans) and thermodynamic-based algorithms. Analysis of our structural alignment revealed four unusual insertions in the core rRNA structure that are unique to animal 18S rRNA and in general agreement with previously proposed insertion sites for strepsipterans. One curious result is the presence of a large insertion within a hairpin loop of a highly conserved pseudoknot helix in variable region 4. Despite the extraordinary variability in sequence length and composition, this insertion contains the conserved sequences 5'-AUUGGCUUAAA-3' and 5'-GAC-3' that immediately flank a putative helix at the 5'- and 3'-ends, respectively. The longer sequence has the potential to form a nine-basepair helix with a sequence in the variable region 2, consistent with a recent study proposing this tertiary interaction. Our analysis of a larger set of arthropod 18S rRNA sequences has revealed possible errors in some of the previously published strepsipteran 18S rRNA sequences. Thus we find no support for the previously recovered heterogeneity in the 18S molecules of strepsipterans. Our findings lend insight to the evolution of RNA structure and function and the impact large insertions pose on genome size. We also provide a novel alignment template that will improve the phylogenetic placement of the Strepsiptera among other insect taxa.

Introduction

For nearly a decade it has been known that the ribosomal RNA (rRNA) genes of strepsipteran insects possess extraordinarily expanded sequences in less conserved regions of the rRNA molecules when compared to other arthropods (Chalwatzis *et al.*, 1995; Whiting *et al.*, 1997; Hwang *et al.*, 1998; Choe *et al.*, 1999b). This peculiar

characteristic of strepsipteran 18S rRNA has been hypothesized to correlate with the unusual biology exhibited by these bizarre insects (Chalwatzis *et al.*, 1995). However, it has been shown that other organisms with less unusual biologies also have greatly expanded rRNA genes, especially in expansion segments and variable regions (e.g., Schnare *et al.*, 1996; Wuyts *et al.*, 2000; Alvares *et al.*, 2004). In particular, complete sequences of the 18S rRNA gene of several arthropods (three hemipterans and a crustacean) are exceptionally larger than average (1800-1900 bp): 2, 469 bp in the pea aphid, *Acyrtosiphon pisum* (Kwon *et al.*, 1991); 3, 214 bp in the soil bug, *Armadillidium vulgare* (Choe *et al.*, 1999a); 2, 293 in the water flea, *Daphnia pulex* (Crease & Colbourne, 1998); 2, 373 in the California red scale, *Aonidiella aurantii* (Campbell *et al.*, 1994); and 2, 496 in Kellogg's whitefly, *Pealius kelloggii* (Campbell *et al.*, 1995). A list of other unusually long metazoan 18S sequences is provided by Giribet & Wheeler (2001). Given these data, we question whether or not the extremely expanded rRNA sequences of strepsipterans can actually be associated with the highly unusual biology exhibited by this bizarre insect taxon.

Indeed, strepsipterans are odd in that the larvae are free living in the first instar, later developing into apodous endoparasites of other insect species (Kathirithamby, 1989). Females (except Mengenillidae) reside inside their hosts for the remainder of their life. The majority of the male life cycle is spent as a larval endoparasite, with short-lived winged adults seeking females for reproduction. Particularly strange are the Myrmecolacidae, in which males and females parasitize hosts in different insect orders, a form of parasitism referred to as heterotrophic heteronomy (Walter, 1983).

Myrmecolacid males exploit Hymenoptera (ants) as hosts, and the females parasitize a range of species in several orthopteroid orders (Ogloblin, 1939; Kathirithamby, 1991a; Kathirithamby & Hamilton, 1992). While it is generally accepted that koinobiont endoparasites (those living within a mobile and defensive host) have a narrower host range than ectoparasitic ones (Askew & Shaw, 1986; Strand & Peach, 1995), a phenomenon likely due to the constraints of the host immune system on endoparasites (Strand, 1986), strepsipterans defy this rule by having an extremely vast host range relative to species richness. Only 596 species of Strepsiptera have been described as of 2004, yet there have been reports of species parasitizing seven orders and 34 families of Insecta (Kathirithamby, 1989). This wide host range is likely greater than that of any group of parasitoid insects (Kathirithamby *et al.*, 2003) and it is hypothesized that this biology promotes the extreme sexual dimorphism (females are highly reduced morphologically) observed in this insect taxon (Kathirithamby, 1989).

Few entomologists would argue against a correlation between the unusual life history of strepsipterans and weird morphological characteristics. However, evidence for tantamount odd molecular differences is not well known. Perhaps if rampant host switching is associated with an increase in the rate of molecular evolution, then strepsipteran DNA sequences may have undergone an accelerated rate of nucleotide substitution, much like that reported for the stem lineage of Diptera (flies) (Friedrich & Tautz, 1997a, b). However, since no comprehensive molecular phylogeny exists for the order Strepsiptera, as does for the Diptera (Yeates & Wiegmann, 1999; Wiegmann *et al.*, 2003), an accelerated rate of nucleotide evolution in strepsipterans remains a

speculation, particularly due to the paucity of existing rRNA sequences for the order and the difficulty in objectively aligning them with often much shorter sequences from other insects.

Earlier secondary structure models for the strepsipteran rRNAs were predicted from a comparative analysis of a limited number of taxa (Hwang *et al.*, 1998; Choe *et al.*, 1999b). Nevertheless, using three 18S rRNA sequences Choe *et al.* (1999b) identified the locations that contain the majority of the extra length present in the strepsipterans and absent in the other arthropods. We have reevaluated the atypical structure of strepsipteran 18S rRNA with a larger number of available sequences and the prediction of a refined double pseudoknot structure in variable region 4 of 18S rRNA that was published after these earlier predicted models (Wuyts *et al.*, 2000).

In this paper, we report the sequence and secondary structure of the entire 18S rDNA gene from the strepsipteran *Caenocholax fenyesei texensis* Kathirithamby & Johnston 2004 (Myrmecolacidae). Based on our analyses, we: 1. provide a secondary structure model for the entire strepsipteran 18S rRNA that includes the variable regions within the conserved core structure, 2. characterize the higher order ribosome structure in these unique regions of strepsipteran rRNA, 3. offer an alignment with strong covariation support of 175 arthropod sequences that will prove useful for future investigations on arthropod phylogenetics, and 4. discuss how our findings are related to the evolution of genome size in organisms with high occurrences of nucleotide insertions.

Figure 26. The secondary structure model of the nuclear SSU rRNA (18S) from the strepsipteran *Caenocholax fenyasi texensis* (accession number DQ026302). A. Domains I-II. B. Domain III. Helix numbering follows the system of Cannone *et al.* (2002), except for variable region 4 (V4) for which the notation of Wuyts *et al.* (2000) is used. Variable regions are colored light blue and the naming follows Van de Peer *et al.* (1999). Regions colored red depict both highly expanded variable regions and insertions in the core rRNA specific to Strepsiptera. Differences between our sequence and previously published *C. fenyasi* sequences (U65190 and U65191) are colored purple, with insertions (dark arrows), deletions (open arrows) and substitutions (parentheses) shown. Sequences colored green depict conserved motifs within the pseudoknot 13/14 insertion. Helices aligned across all sampled panarthropods are boxed in grey. Regions of ambiguous alignment are boxed in green and characterized following the method of Gillespie (2004). A single ambiguity in helix **H829a** is boxed in red. Base-pairing (where there is strong comparative support) and base triples are shown connected by continuous lines. Base-pairing is indicated as follows: standard canonical pairs by lines (C-G, G-C, A-U, U-A); wobble G•U pairs by dots (G•U); A•G pairs by open circles (A^oG); other non-canonical pairs by filled circles (e.g. C•A). Universal primers, as well as primers designed in this study, are mapped on the structure in orange with the first primer position circled. A primer table is posted at the jRNA website. Diagram was generated using the program XRNA (Weiser, B. & Noller, H., University of California at Santa Cruz) with severe manual adjustment.

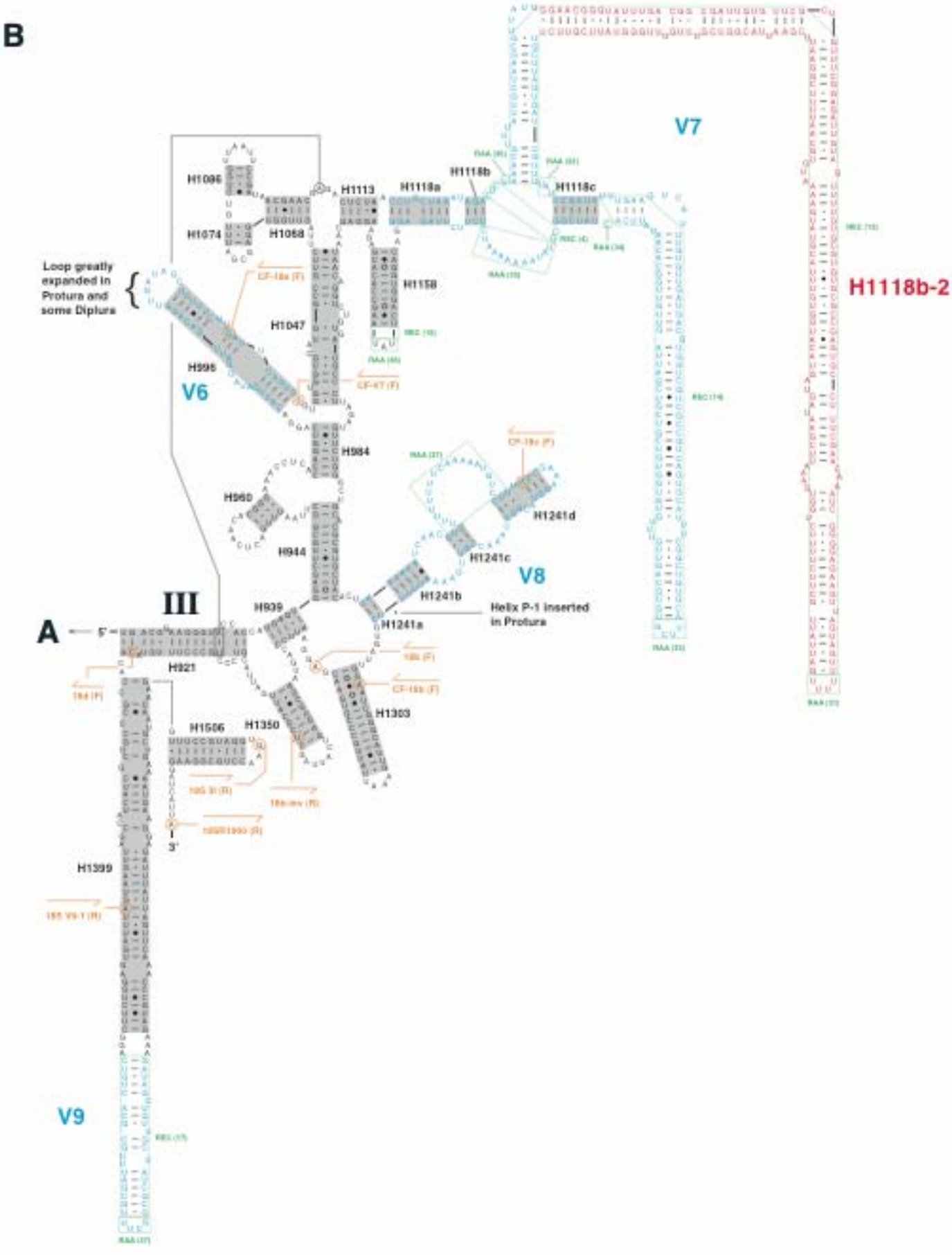


Figure 26 Continued.

Results and discussion

Predicted secondary structure

Our predicted 18S rRNA secondary structure for *C. f. texensis* is shown in Figure 26.

The diagram follows the secondary structural model of *Drosophila melanogaster* (Gutell, 1993; 1994; Cannone *et al.*, 2002), with refinement to the variable region 4 (Van de Peer *et al.*, 1999) made from the model of Wuyts *et al.* (2000). The length of the 18S rRNA in *C. f. texensis* is 2,864 nts, which is currently the fourth largest arthropod 18S rRNA known (only the strepsipterans *Xenos vesparum* and *X. pecki* of the family Stylopidae and the soil bug *Armadillidium vulgare* (Crustacea) are larger). While the core rRNA sequences of strepsipterans superimpose onto the other arthropod sequences and predicted structure with little or no ambiguity, most of the variable regions previously characterized for SSU rRNA are greatly expanded (Table 23). Additionally, several regions of strepsipteran 18S rRNA, localized within universally conserved core elements, contain unique features of animal SSU rRNA (Choe *et al.*, 1999b). These characteristics of strepsipteran nuclear SSU rRNA are discussed below.

18S rRNA features unique to Strepsiptera

H143: This putative helix is present in some arthropods as a small helix (2-4 bps) except for the psocopteran *Liposcelis* sp. that has a putative 12 bp helix that is not energetically stable. In contrast this helix in strepsipterans is highly expanded, ranging from 15 to 19 bps. Comparative support for this helix is minimal across closely related

Table 23. Summary statistics on the variable regions of the 18S rRNA in the major arthropod classes and hexapod orders¹. Within columns, A/U composition is listed followed by ranges of nucleotide lengths within defined variable regions. See Figure 1 for location of variable regions.

	V1	H143	V2	H184c	H198b	V3
Tardigrada ⁵	.55, 33	.73, 11	.61, 100	0, 0	.40, 5	.44, 34
Onychophora	.52, 37-38	.40, 12-13	.41, 97-118	0, 0	.71, 3-4	.46, 35
Chelicerata	.53, 33-35	.66, 10-12	.48, 95-128	0, 0	.42, 5-6	.48, 34
Myriapoda	.51, 33-42	.62, 10	.45, 99-101	0, 0	.52, 5-19	.44, 34-36
Crustacea	.45, 33	.63, 9-13	.50, 90-116	0, 0	.27, 4-5	.40, 34
Collembola	.61, 33-34	.69, 9	.53, 92-94	0, 0	.32, 4-5	.48, 34
Protura	.42, 34-35	.70, 10	.44, 130-132	0, 0	.56, 5-6	.29, 31
Diplura	.32, 37	.42, 10	.38, 99-165	0, 0	.43, 8-10	.16, 31-35
Archaeognatha	.49, 33-34	.67, 10-12	.49, 95-114	0, 0	.31, 5-6	.43, 34
Thysanura	.56, 33-36	.74, 11	.58, 102-117	0, 0	.56, 6	.38, 34
Odonata	.49, 32-33	.61, 11-12	.50, 94-103	0, 0	.25, 5-6	.38, 34
Ephemeroptera	.52, 29-33	.70, 10-11	.46, 102-105	0, 0	.34, 5-6	.38, 34
Dermaptera	.53, 33	.83, 12	.68, 103	0, 0	.54, 6	.38, 34
Plecoptera	.54, 30-33	.66, 11-12	.53, 90-123	0, 0	.33, 5-6	.38, 34
Grylloblattodea	.48, 33	.67, 12	.51, 128-130	0, 0	.40, 5	.38, 34
Embioptera	.51, 30-33	.58, 12	.43, 102	0, 0	.33, 6	.38, 34
Blattaria	.47, 33	.64, 10	.48, 98-133	0, 0	.38, 5-6	.38, 34
Isoptera	.48, 33	.61, 9-10	.48, 101-133	0, 0	.33, 6	.38, 34
Mantodea	.48, 33	.60, 10	.49, 97	0, 0	.17, 6	.38, 34
Phasmatodea	.53, 33	.65, 10	.53, 104-105	0, 0	.25, 6	.40, 34
Orthoptera	.51, 33	.58, 10-11	.51, 104-105	0, 0	.44, 6	.38, 34
Hemiptera	.53, 33	.57, 10-12	.49, 99-107	0, 0	.47, 6	.39, 34
Psocoptera	.53, 33-48	.49, 10-39	.55, 102-146	0, 0	.50, 5-6	.46, 35
Phthiraptera	.52, 33	.64, 12-16	.53, 224-244	0, 0	.30, 5-7	.48, 35-36
Thysanoptera ⁵	.55, 33	.73, 11	.51, 102	0, 0	.33, 6	.38, 34
Hymenoptera	.47, 33	.67, 12	.56, 110-117	0, 0	.42, 6	.38, 34
Mecoptera	.49, 33	.64, 12	.58, 109-110	0, 0	.47, 5	.38, 34
Siphonaptera ⁵	.48, 33	.64, 11	.55, 110	0, 0	.40, 5	.38, 34
Lepidoptera	.43, 33	.65, 10	.55, 99-110	0, 0	.30, 5	.44, 34
Trichoptera	.51, 33	.65, 10	.46, 114	0, 0	.40, 5	.35, 34
Neuroptera	.47, 33	.61, 11-12	.50, 119-124	0, 0	.40, 5	.38, 34
Coleoptera	.49, 33	.67, 12	.52, 103-119	0, 0	.29, 4-5	.38, 34
Diptera	.61, 33	.82, 10-12	.65, 88-110	0, 0	.50, 5	.57, 35-36
Strepsiptera	.64, 38-69*	.72, 49-57***	.66, 149-225***	.77, 54,-194***	.58, 32-71**	.53, 34

Table 23 Continued.

	V4					
	E23-1, 2 ^b	E23-5 ^c	E23-6	E23-7 ^d	E23-8-14 ^e	total
Tardigrada ^s	.52, 56	0, 0	0, 0	.49, 41	.61, 126	.56, 242
Onychophora	.17, 63-64	0, 0	.25, 8	.30, 38-43	.51, 132-147	.39, 255-274
Chelicerata	.38, 51-55	.41, 2-30	.43, 7	.40, 33-36	.61, 125-126	.50, 231-365
Myriapoda	.33, 54-61	1, 1	.36, 11	.35, 35-40	.54, 127-137	.49, 236-265
Crustacea	.36, 55-74	.40, 5	.25, 8	.39, 27-40	.59, 125-133	.48, 233-278
Collembola	.46, 54-55	0, 0	0, 0	.54, 33-35	.60, 125	.55, 231-234
Protura	.28, 73	.27, 61-64	0, 0	.27, 11	.50, 130	.38, 292-295
Diplura	.28, 58-65	----- ^f	----- ^g	----- ^h	.46, 123-131	.37, 235-524
Archaeognatha	.33, 54-66	0, 0	0, 0	.41, 35-36	.57, 124-130	.47, 234-252
Thysanura	.37, 55-59	0, 0	0, 0	.52, 35-38	.61, 127-131	.52, 201-245
Odonata	.32, 57-58	.31, 16-19	.73, 5	.46, 35	.59, 129	.47, 261-265
Ephemeroptera	.33, 56-58	.21, 21-43	.62, 5-9	.33, 33-40	.60, 129-131	.45, 265-289
Dermaptera	.52, 58-60	.59, 68-85	.70, 5	.46, 28-41	.64, 126	.58, 310-331
Plecoptera	.34, 64-68	.30, 54-237 ^k	.58, 5-7	.45, 34-39	.63, 126-129	.44, 303-488
Grylloblattodea	.43, 63	.50, 246-250	.86, 7	.43, 36	.60, 128	.52, 499-503
Embiopoda	.35, 62	.30, 37	.89, 9	.43, 34	.59, 128	.47, 289
Blattaria	.32, 59-62	.23, 20-115	.81, 8	.37, 32-37	.58, 127-130	.42, 265-359
Isoptera	.32, 61-62	.22, 70-73	.74, 7	.36, 35-39	.57, 130-132	.41, 251-326
Mantodea	.35, 62	.19, 20-45	.86, 7	.38, 35	.58, 131	.44, 273-298
Phasmatodea	.39, 62	.24, 44	.56, 9	.37, 34	.61, 129	.46, 297
Orthoptera	.38, 62-64	.24, 60-87	.71, 7-9	.37, 34-35	.60, 127-129	.44, 312-342
Hemiptera	.34, 58-60	.27, 66-79	.53, 7-9	.40, 34-40	.61, 129-132	.44, 314-329
Psocoptera	.43, 61	.37, 34-53	.58, 9-10	.36, 34-35	.63, 128-131	.50, 285-309
Phthiraptera	.35, 60-62	.32, 56 ^l	.61, 6-12	.32, 34-39	.57, 128-134	.44, 309-464
Thysanoptera ^s	.33, 60	.27, 41	.56, 18	.38, 34	.61, 126	.46, 298
Hymenoptera	.42, 58-60	.30, 25	.88, 8	.41, 35	.62, 129	.52, 274-276
Mecoptera	.41, 59-62	.36, 26-27	.81, 9	.44, 35	.65, 130	.54, 278-282
Siphonaptera ^s	.38, 60	.33, 27	.78, 9	.37, 35	.63, 128	.51, 278
Lepidoptera	.34, 60	.35, 25-26	.84, 9-10	.37, 35	.66, 121-132	.51, 270-281
Trichoptera	.38, 33-58	.25, 18-26	.41, 9-18	.24, 25-29	.61, 132-134	.47, 253-267
Neuroptera	.38, 60-61	.25, 119-161	.81, 8-13	.36, 34-35	.61, 131	.42, 377-414
Coleoptera	.29, 58-60	.27, 25-104	.64, 7-13	.32, 34-35	.62, 131	.45, 275-360
Diptera	.55, 62-78	.69, 19-54	.83, 9-33	.54, 33-38	.67, 129-131	.63, 264-336
Strepsiptera	.67, 67-74	.70, 40-140	.75, 90-217***	.60, 32-146*	.71, 259-505***	.70, 599-859***

Table 23 Continued.

	V5	V6	H1118b-2	V7 H1118c	total	V8	V9
Tardigrada ^s	.50, 44	.57, 44	0, 0	.43, 21	.56, 59	.55, 71	.44, 41
Onychophora	.27, 44-46	.47, 44	0, 0	.53, 21-26	.47, 62-65	.35, 64-66	.28, 41-67
Chelicerata	.37, 44	.58, 44	0, 0	.44, 17-26	.52, 49-66	.43, 61-65	.22, 41-43
Myriapoda	.36, 44-45	.53, 44	0, 0	.40, 18-32	.46, 52-64	.42, 62-67	.24, 28-49
Crustacea	.38, 44-45	.54, 44	0, 0	.46, 17-24	.51, 49-62	.40, 58-63	.30, 38-67
Collembola	.46, 44-45	.52, 44	0, 0	.57, 17-20	.55, 51-53	.52, 58-59	.39, 41
Protura	.26, 45	.54, 52	0, 0	.47, 25-26	.45, 58-60	.35, 112	.25, 56-59
Diplura	.28, 44-51	.46, 44-51	0, 0	.16, 20-22	.28, 52-54	.33, 62-64	.32, 38-49
Archaeognatha	.37, 43-44	.56, 44	0, 0	.37, 22-27	.46, 52-58	.38, 61	.20, 50-59
Thysanura	.37, 44	.59, 44	0, 0	.72, 15-22	.61, 57-60	.43, 59-63	.42, 35-47
Odonata	.40, 43-44	.54, 44	0, 0	.61, 22	.53, 56-57	.39, 59	.26, 47-50
Ephemeroptera	.34, 44	.57, 44	0, 0	.37, 21-22	.46, 54-56	.39, 59-62	.25, 39-44
Dermaptera	.40, 44	.54, 44	0, 0	.43, 20	.51, 61	.42, 59	.52, 44-45
Plecoptera	.33, 43-44	.59, 44	1, 5	.42, 25	.52, 69-74	.43, 61-63	.30, 39-44
Grylloblattodea	.32, 44	.52, 44	0, 0	.35, 23	.53, 64	.44, 59	.43, 40
Embioptera	.34, 44	.52, 44	0, 0	.39, 22	.49, 61	.39, 59	.38, 41
Blattaria	.34, 44	.51, 42-44	0, 0	.37, 22-23	.48, 50-63	.39, 59	.29, 38-42
Isoptera	.34, 44	.50, 43-44	0, 0	.35, 22	.49, 50-62	.40, 59	.18, 38
Mantodea	.34, 44	.50, 44	0, 0	.32, 22	.44, 62	.39, 59	.21, 41
Phasmatodea	.34, 44	.52, 44	0, 0	.35, 23	.49, 63	.41, 59	.35, 41
Orthoptera	.31, 44	.54, 44	0, 0	.41, 21-23	.50, 60-63	.42, 59	.34, 41-44
Hemiptera	.34, 44-48	.56, 44	0, 0	.41, 19-25	.53, 60-67	.39, 59-62	.30, 34-44
Psocoptera	.38, 44-46	.63, 43-44	.41, 9-32	.38, 19-21	.49, 66-95	.46, 63-64	.43, 45-46
Phthiraptera	.42, 44-45	.54, 44	.69, 9-28	.42, 18-104	.53, 78-137	.47, 63-65	.34, 45-73
Thysanoptera ^s	.43, 44	.57, 44	0, 0	.57, 23	.51, 63	.41, 59	.30, 40
Hymenoptera	.35, 44	.53, 44	.90, 5	.43, 49	.52, 94	.41, 59	.28, 40
Mecoptera	.37, 44	.57, 44	.80, 5	.45, 62-67	.45, 105-112	.45, 59	.43, 40-41
Siphonaptera ^s	.34, 44	.52, 44	.80, 5	.41, 63	.49, 108	.44, 59	.35, 40
Lepidoptera	.39, 43-44	.55, 44	0, 0	.38, 47	.48, 85	.42, 59	.26, 38
Trichoptera	.47, 44-45	.52, 44	0, 0	.34, 80	.45, 125	.39, 59	.30, 40
Neuroptera	.41, 44	.52, 44	.78, 9	.30, 155-215	.37, 205-264	.43, 59	.41, 41
Coleoptera	.32, 44	.52, 44	.75, 8	.29, 27-136	.40, 62-184	.37, 59	.34, 40-45
Diptera	.54, 43-45	.66, 44	.42, 8-11	.59, 21-103	.63, 68-164	.52, 57-60	.55, 41-52
Strepsiptera	.48, 45-46	.60, 44	.63, 143-307***	.59, 128-300***	.61, 432-569***	.60, 60-72	.54, 41-43

Table 23 Continued.

	V5	V6	H1118b-2	V7 H1118c	total	V8	V9
Tardigrada ⁵	.50, 44	.57, 44	0, 0	.43, 21	.56, 59	.55, 71	.44, 41
Onychophora	.27, 44-46	.47, 44	0, 0	.53, 21-26	.47, 62-65	.35, 64-66	.28, 41-67
Chelicerata	.37, 44	.58, 44	0, 0	.44, 17-26	.52, 49-66	.43, 61-65	.22, 41-43
Myriapoda	.36, 44-45	.53, 44	0, 0	.40, 18-32	.46, 52-64	.42, 62-67	.24, 28-49
Crustacea	.38, 44-45	.54, 44	0, 0	.46, 17-24	.51, 49-62	.40, 58-63	.30, 38-67
Collembola	.46, 44-45	.52, 44	0, 0	.57, 17-20	.55, 51-53	.52, 58-59	.39, 41
Protura	.26, 45	.54, 52	0, 0	.47, 25-26	.45, 58-60	.35, 112	.25, 56-59
Diplura	.28, 44-51	.46, 44-51	0, 0	.16, 20-22	.28, 52-54	.33, 62-64	.32, 38-49
Archaeognatha	.37, 43-44	.56, 44	0, 0	.37, 22-27	.46, 52-58	.38, 61	.20, 50-59
Thysanura	.37, 44	.59, 44	0, 0	.72, 15-22	.61, 57-60	.43, 59-63	.42, 35-47
Odonata	.40, 43-44	.54, 44	0, 0	.61, 22	.53, 56-57	.39, 59	.26, 47-50
Ephemeroptera	.34, 44	.57, 44	0, 0	.37, 21-22	.46, 54-56	.39, 59-62	.25, 39-44
Dermaptera	.40, 44	.54, 44	0, 0	.43, 20	.51, 61	.42, 59	.52, 44-45
Plecoptera	.33, 43-44	.59, 44	1, 5	.42, 25	.52, 69-74	.43, 61-63	.30, 39-44
Grylloblattodea	.32, 44	.52, 44	0, 0	.35, 23	.53, 64	.44, 59	.43, 40
Embioptera	.34, 44	.52, 44	0, 0	.39, 22	.49, 61	.39, 59	.38, 41
Blattaria	.34, 44	.51, 42-44	0, 0	.37, 22-23	.48, 50-63	.39, 59	.29, 38-42
Isoptera	.34, 44	.50, 43-44	0, 0	.35, 22	.49, 50-62	.40, 59	.18, 38
Mantodea	.34, 44	.50, 44	0, 0	.32, 22	.44, 62	.39, 59	.21, 41
Phasmatodea	.34, 44	.52, 44	0, 0	.35, 23	.49, 63	.41, 59	.35, 41
Orthoptera	.31, 44	.54, 44	0, 0	.41, 21-23	.50, 60-63	.42, 59	.34, 41-44
Hemiptera	.34, 44-48	.56, 44	0, 0	.41, 19-25	.53, 60-67	.39, 59-62	.30, 34-44
Psocoptera	.38, 44-46	.63, 43-44	.41, 9-32	.38, 19-21	.49, 66-95	.46, 63-64	.43, 45-46
Phthiraptera	.42, 44-45	.54, 44	.69, 9-28	.42, 18-104	.53, 78-137	.47, 63-65	.34, 45-73
Thysanoptera ⁵	.43, 44	.57, 44	0, 0	.57, 23	.51, 63	.41, 59	.30, 40
Hymenoptera	.35, 44	.53, 44	.90, 5	.43, 49	.52, 94	.41, 59	.28, 40
Mecoptera	.37, 44	.57, 44	.80, 5	.45, 62-67	.45, 105-112	.45, 59	.43, 40-41
Siphonaptera ⁵	.34, 44	.52, 44	.80, 5	.41, 63	.49, 108	.44, 59	.35, 40
Lepidoptera	.39, 43-44	.55, 44	0, 0	.38, 47	.48, 85	.42, 59	.26, 38
Trichoptera	.47, 44-45	.52, 44	0, 0	.34, 80	.45, 125	.39, 59	.30, 40
Neuroptera	.41, 44	.52, 44	.78, 9	.30, 155-215	.37, 205-264	.43, 59	.41, 41
Coleoptera	.32, 44	.52, 44	.75, 8	.29, 27-136	.40, 62-184	.37, 59	.34, 40-45
Diptera	.54, 43-45	.66, 44	.42, 8-11	.59, 21-103	.63, 68-164	.52, 57-60	.55, 41-52
Strepsiptera	.48, 45-46	.60, 44	.63, 143-307***	.59, 128-300***	.61, 432-569***	.60, 60-72	.54, 41-43

Table 23 Continued.

- ^a Numbers have been adjusted to exclude taxa wherein sequences are partial for a given region. Bolded values depict uniquely long sequence lengths in the Strepsiptera. The mean range of strepsipterans is significantly greater than that of all other orders (PROC GLM Sheffe $p < .01$; SAS 8.22; SAS Inst., Inc. Cary, NC). For each region the ranges observed for strepsipterans were compared to that of all other orders using orthogonal contrasts in PROC GLM of SAS (*** $p < .0001$, ** $p < .01$, * $p < .05$)
- ^b Includes RAA (14)
- ^c Includes RAA (15) and RAA (17)
- ^d Includes RAA (19) and RAA (21)
- ^e Includes RAA (22)
- ^f Only one taxon sampled
- ^g Structures predicted in this region (37-310 nts) did not conform to the alignment model
- ^h Two sequences, *Isoperla obscura* (196 nts) and *Megarocyis stigmata* (231 nts) have sequences with odd, unalignable structures
- ⁱ Three taxa, *Pediculus humanus*, *Haematomyzus elephantis* and *Heterodoxus calabyi*, range from 128-211 nts and form odd structures

taxa; however, the helix predicted with a thermodynamic-based algorithm is similar in all strepsipterans (Fig. 27). This insertion site was proposed by Choe *et al.* (1999b) to foster a helix ranging from 15 to 17 bps, yet their proposed structures lack from one to four basepairs in the basal region of the helix, with three alternate and less stable basepairs in one taxon (*Xenos vesparum*, X77784). These inconsistencies resulted from our model being based on the boundaries of core helices **H122** and **H144** of the *E. coli* 16S-like model (Cannone *et al.*, 2002), which has been verified by the recent crystalline structures of the ribosome (Ban *et al.*, 2000; Wimberly *et al.*, 2000; Schlueder *et al.*, 2000; Yusupov *et al.*, 2001; Gutell *et al.*, 2002). Interestingly, the pogonophoran worm *Siboglinum fiordicum* also contains a helical insertion in this region of the 18S (Winnepeenninckx *et al.*, 1995).

H184c: This helix occurs strictly in Strepsiptera and ranges from 21 bps to 73 bps (Table 23). The majority of length variation in this region of the 18S in non-strepsipteran insects occurs to the 5'-side of helix **H184c** (RAA (8)), with the 3'-sequence just before helix **H198** unambiguously-aligned across Arthropoda. Adjacent to the 5'-end of this conserved sequence (flanking the 3'-end of helix **H184c**) occurs a 5'-AA-3' sequence found only in strepsipterans (Fig. 27). The predicted structure for *X. vesparum* by Choe *et al.* (1999b) for **H184c** differs from our model and is likely based on discrepancies in algorithms used to predict both structures. However, the predictions by Choe *et al.* (1999b) for *Stylops melittae* (X89440, Stylopidae) and *Mengenilla chobauti* (X89441, Mengenillidae) of **H184c** are inaccurate due to the inclusion of some paired nucleotides in **H198b** within their structures (Fig. 28).

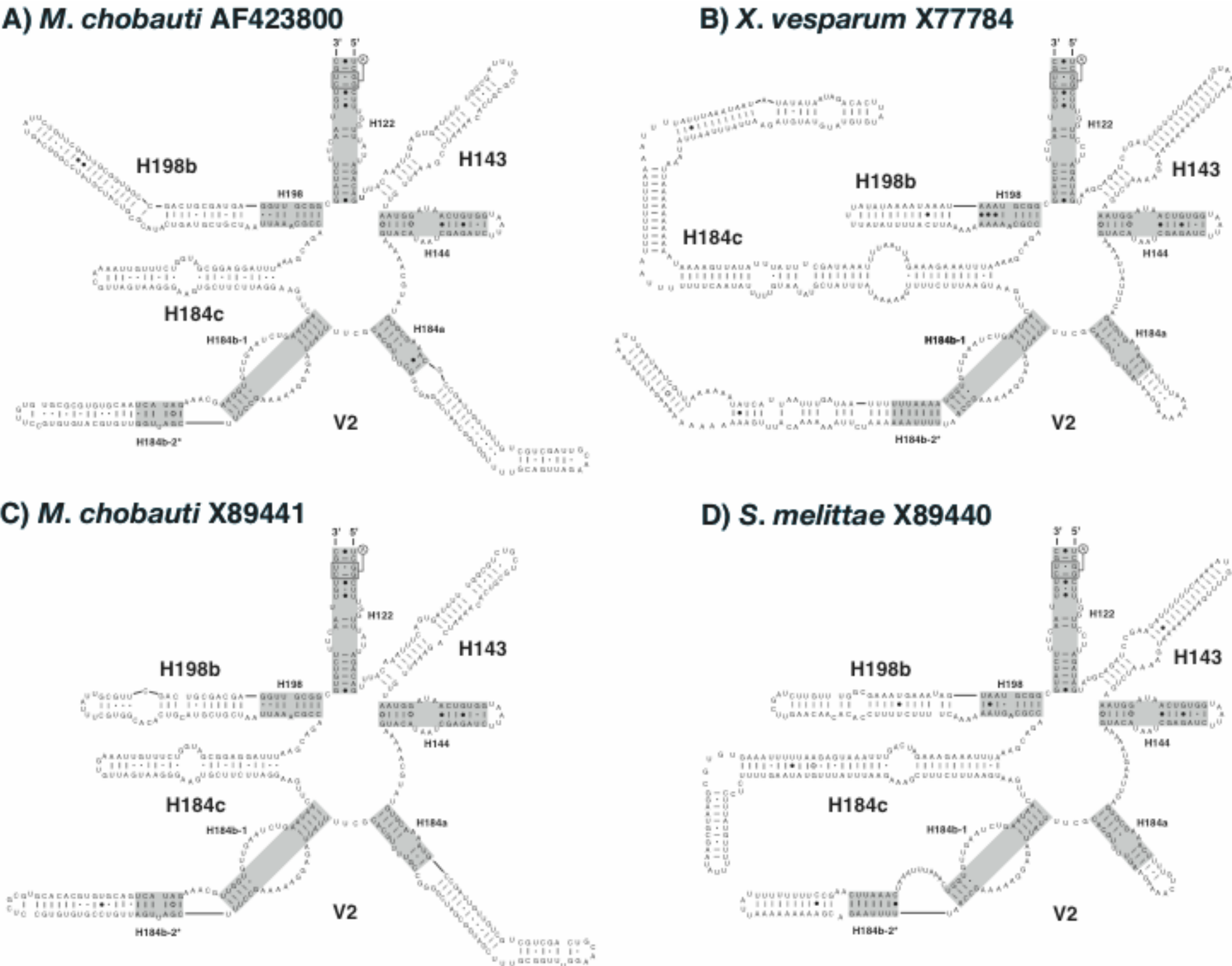
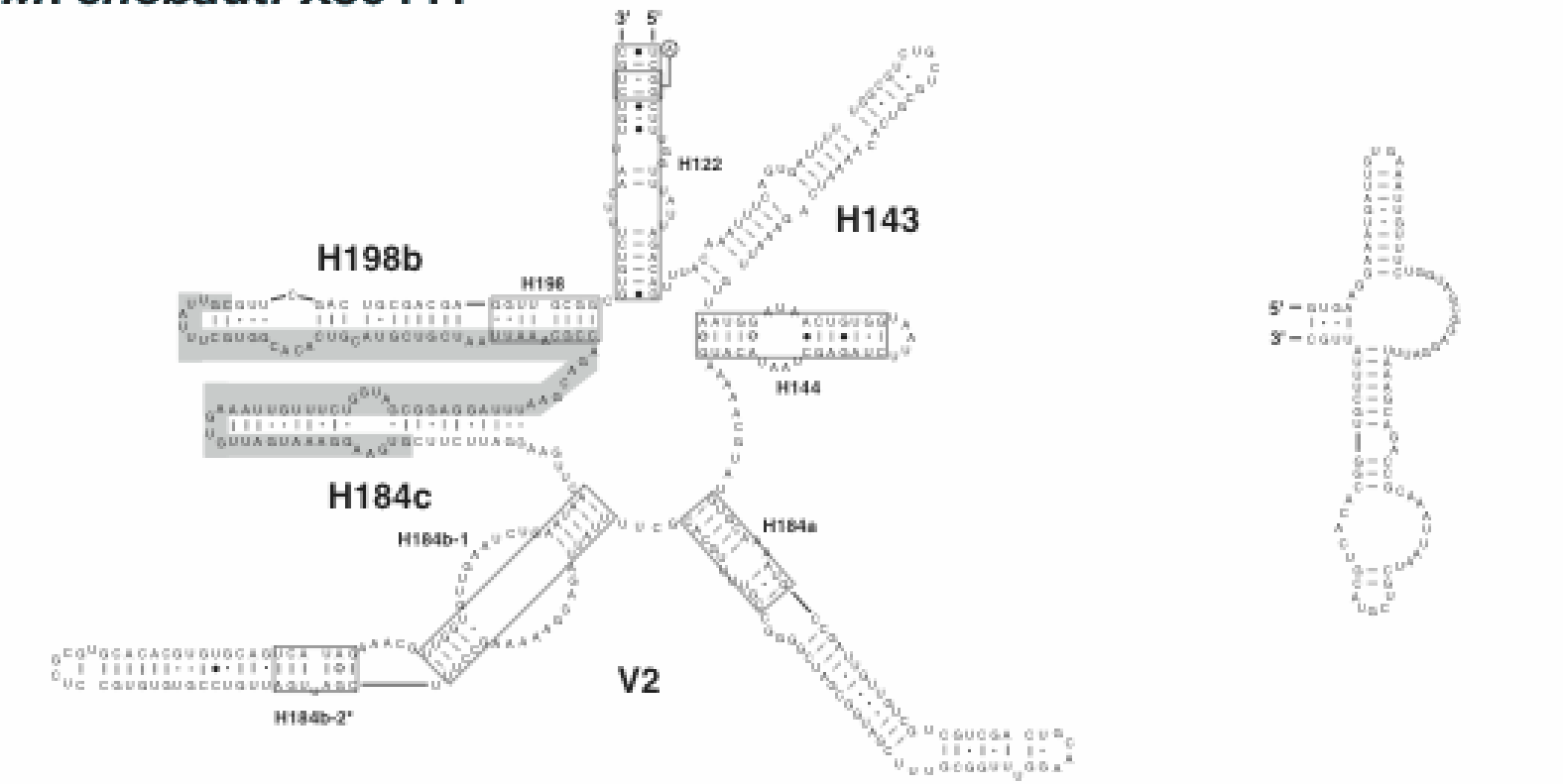


Figure 27. A gallery of diverse secondary structure diagrams of the variable region 2 (V2) and related core elements from selected strepsipterans. Helices **H143**, **H184c** and **H198b** are specific to Strepsiptera. The explanations of base-pair symbols, helix numbering and reference for software used to construct structure diagrams are in Figure 26.

M. chobauti X89441



S. melittae X89440

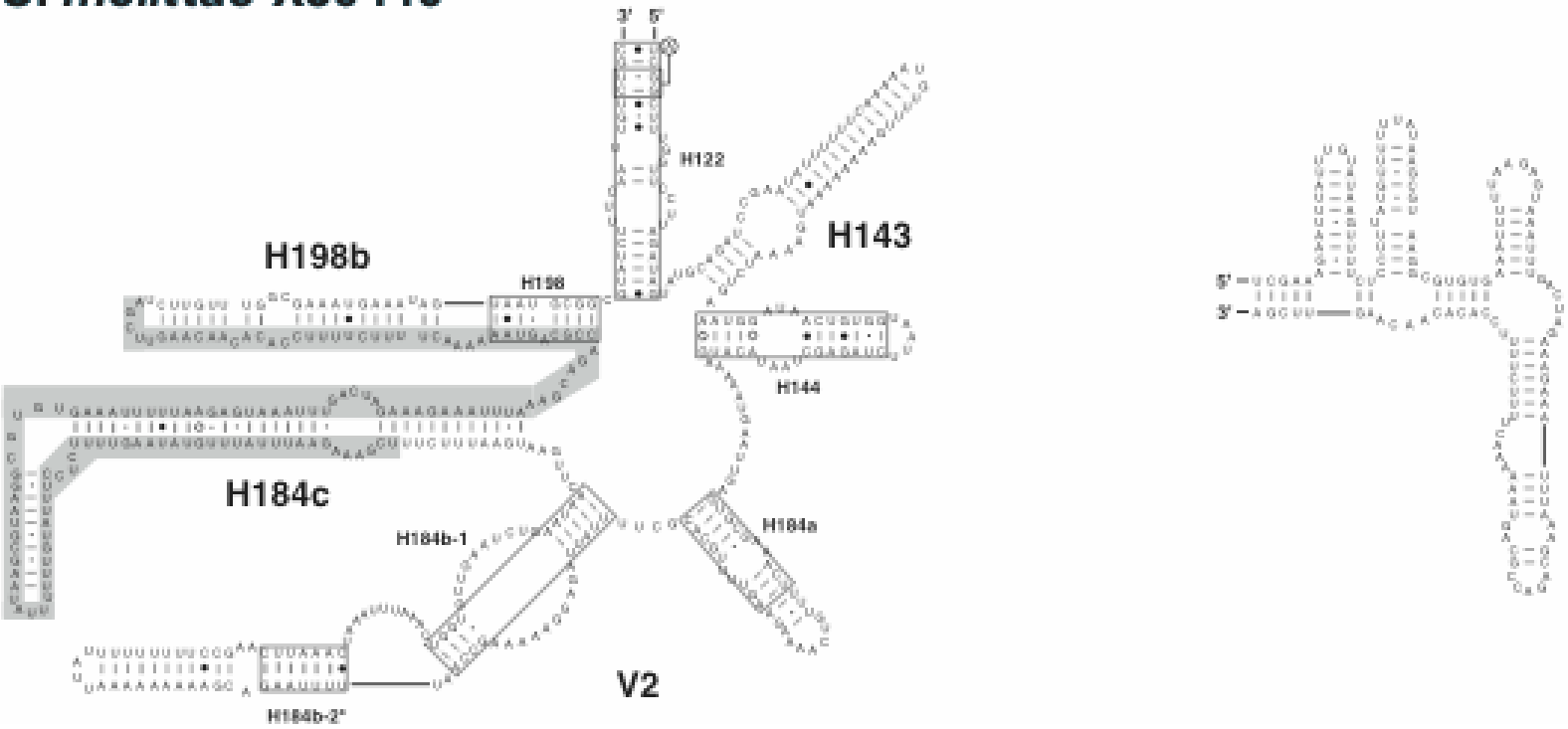


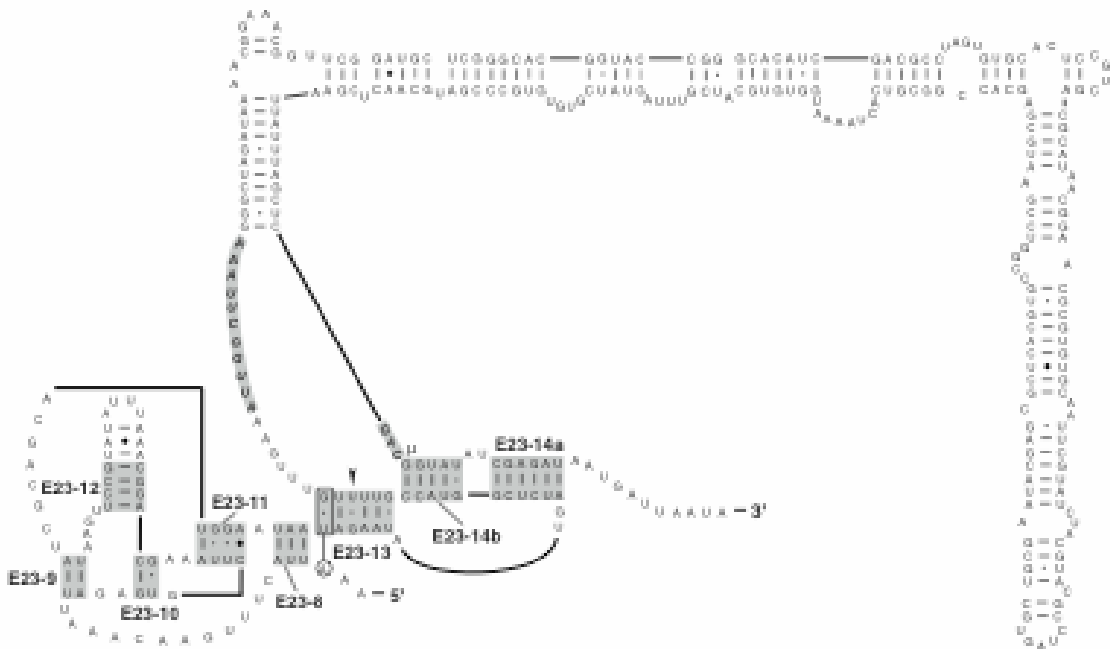
Figure 28. A comparison between our predicted structures for helix **H184c** and those of Choe *et al.* (1999b). Shaded regions in our diagrams (left) depict nucleotides found within their models. The explanations of base-pair symbols, helix numbering and reference for software used to construct structure diagrams are in Figure 26. Models from Choe *et al.* (1999b) were reproduced manually.

H198b: The extension of helix **H198** (8 bps) occurs in several arthropod groups and is usually no greater than 2 bps. However, this helix is greatly expanded in Strepsiptera and ranges from 21 to 34 bps (Fig. 27). In contrast to variable regions (Gerbi, 1985), it is unclear how the expansion and contraction of an otherwise highly conserved core helix affects ribosome assembly and function. Similar evidence for the expansion and contraction of a conserved core helix has recently been detected in helix **H604** in domain II of 28S rRNA in Hymenoptera (Gillespie *et al.*, 2005a, b). Interestingly, unpublished data from our labs suggests helix **H604** is extraordinarily hyper-variable in sequence length and base composition in strepsipterans.

E23-13/14: A large and unusual insertion occurs exclusively in Strepsiptera in the hairpin-stem loop of the pseudoknot 13/14 in the V4 region (Choe *et al.*, 1999b). Inserts vary from 118 nts to 366 nts, with putative helical regions supported by thermodynamic algorithms and comparative evidence across closely related taxa. A gallery of diverse secondary structure predictions illustrates the lack of sequence conservation and structure in this insertion site (Fig. 29). Despite the lack of significant conservation in this region across the strepsipterans, the conserved sequence 5'-AUUGGCUUAAA-3' always occurs immediately 5' to a helical structure that is flanked on its 3'-end by a 5'-GAC-3' sequence. However, the precise location of this conserved sequence does vary within this insertion (Fig. 29). Interestingly, these two highly conserved sequences only differ in the two individuals of Mengenillidae, the proposed sister taxon to the rest of the families of Strepsiptera (Kinzelbach, 1971; Kathirithamby, 1989). The distribution of these hyper-variable sequences within the two conserved

Figure 29. A gallery of diverse secondary structure diagrams of the insertion within the hairpin loop of pseudoknot 13/14 within variable region 4 (V4) from selected strepsipterans. The conserved sequences described in the text are bold-italicized and shaded. Differences between *C. fenyesei* (U65190) and *C. fenyesei* (U65191), and between *C. fenyesei* (U65190) and *T. mexicana* (U65159) are depicted using the same symbols described in Figure 1, with μ representing a substitution. F. The "short" unpublished sequence of *X. pecki* from Whiting *et al.* (1997) is shown under *X. pecki* (U65164) with an asterisk depicting the missing sequence and potential inclusion of a putative cloning vector. The explanations of base-pair symbols, helix numbering and reference for software used to construct structure diagrams are in Figure 26.

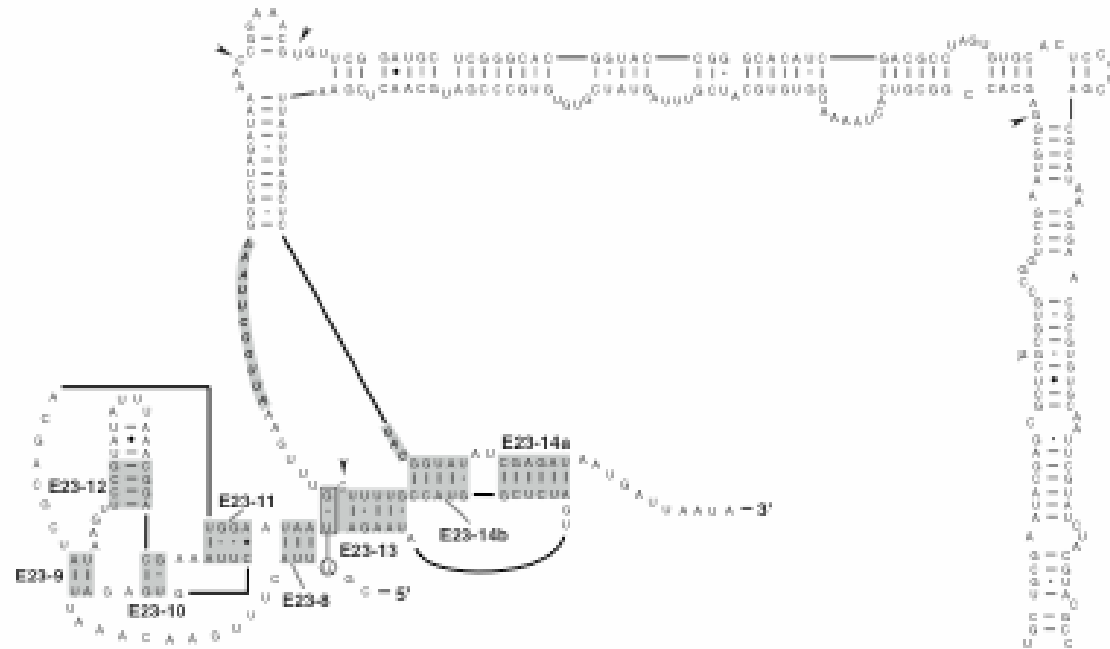
A) *C. fenyesi* U65160



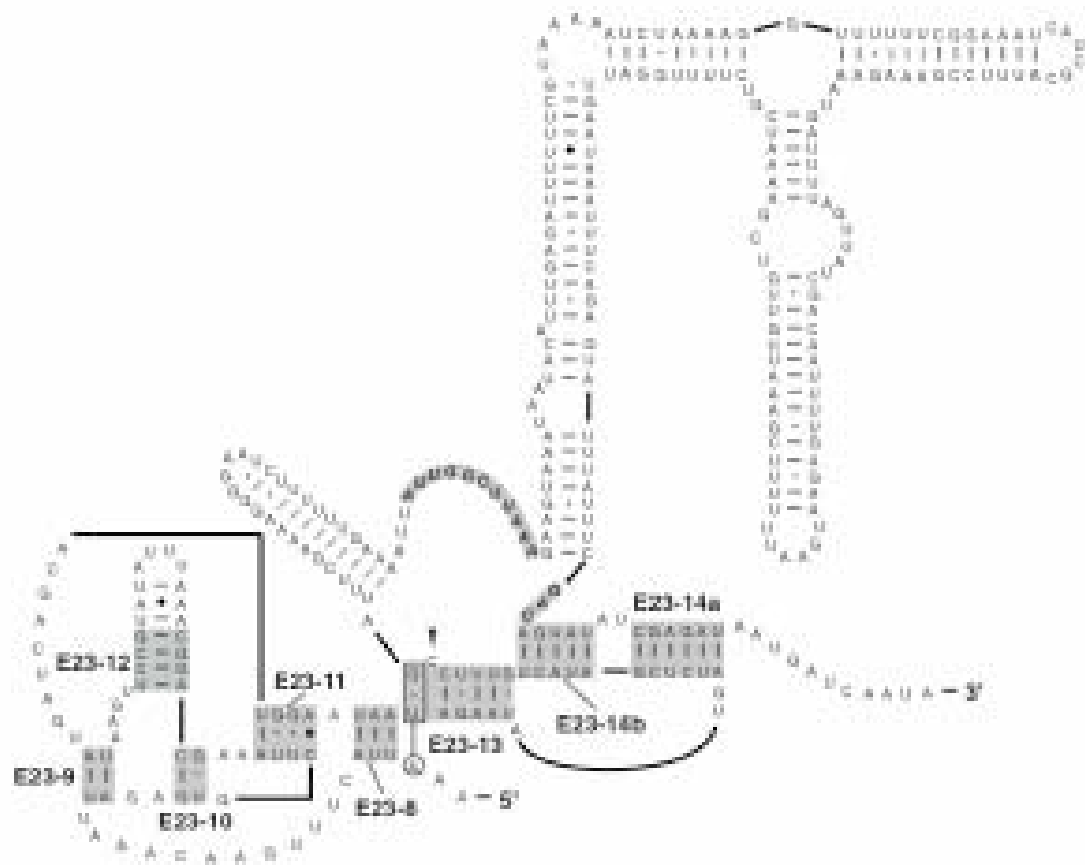
B) *C. fenyesi* U65161



C) *T. mexicana* U65159



D) *Crawfordia* sp. U65163



E) *S. melittae* X89440

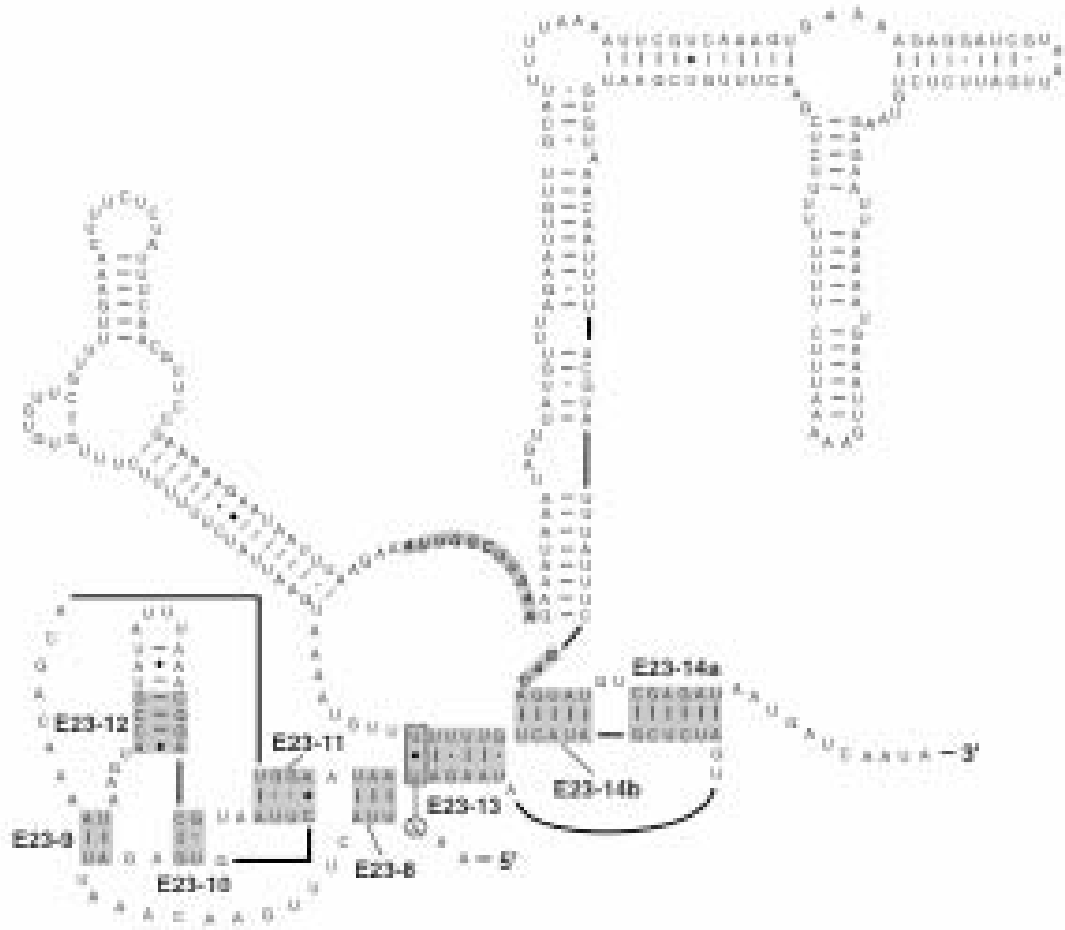
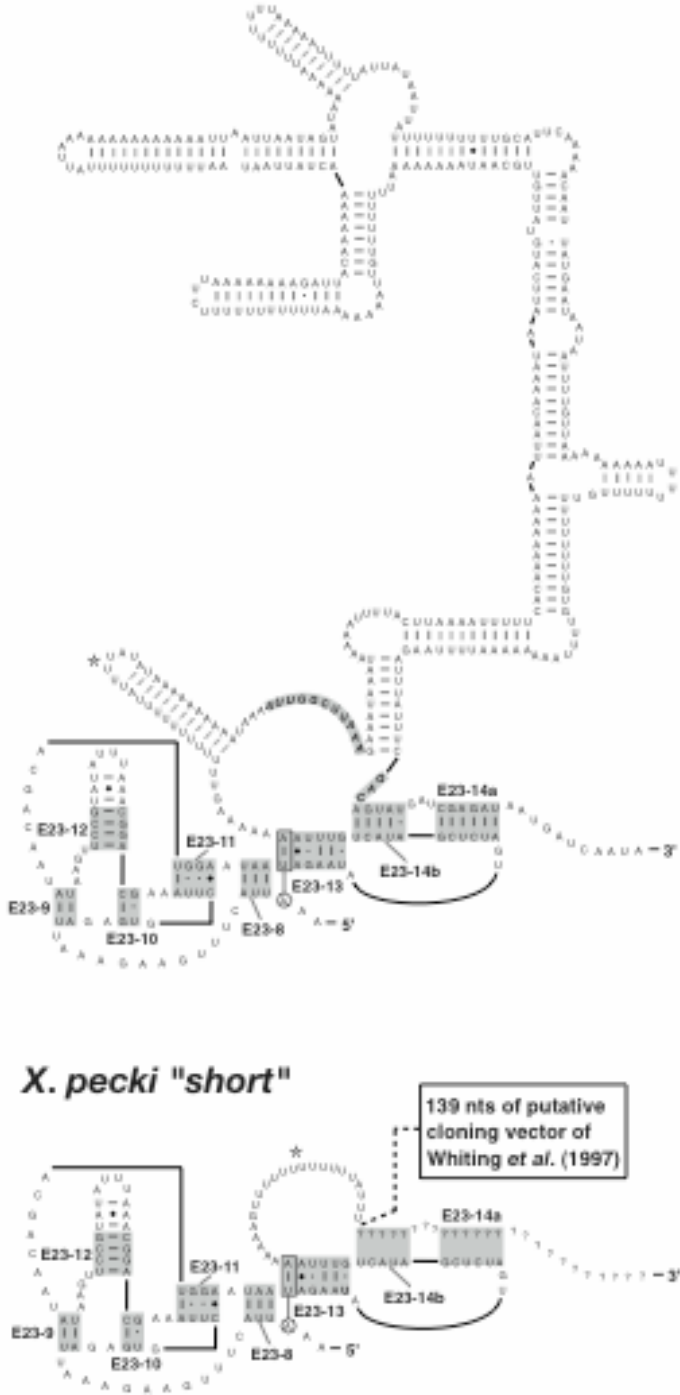
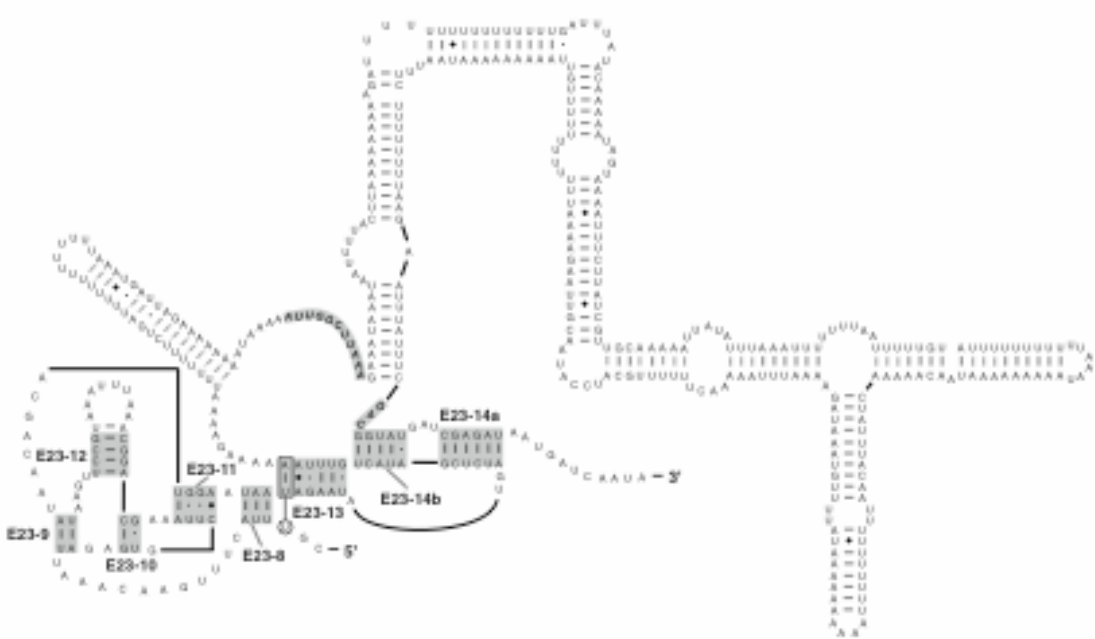


Figure 29 Continued.

F) *X. pecki* U65164



G) *X. vesparum* X77784



H) *X. vesparum* X74763

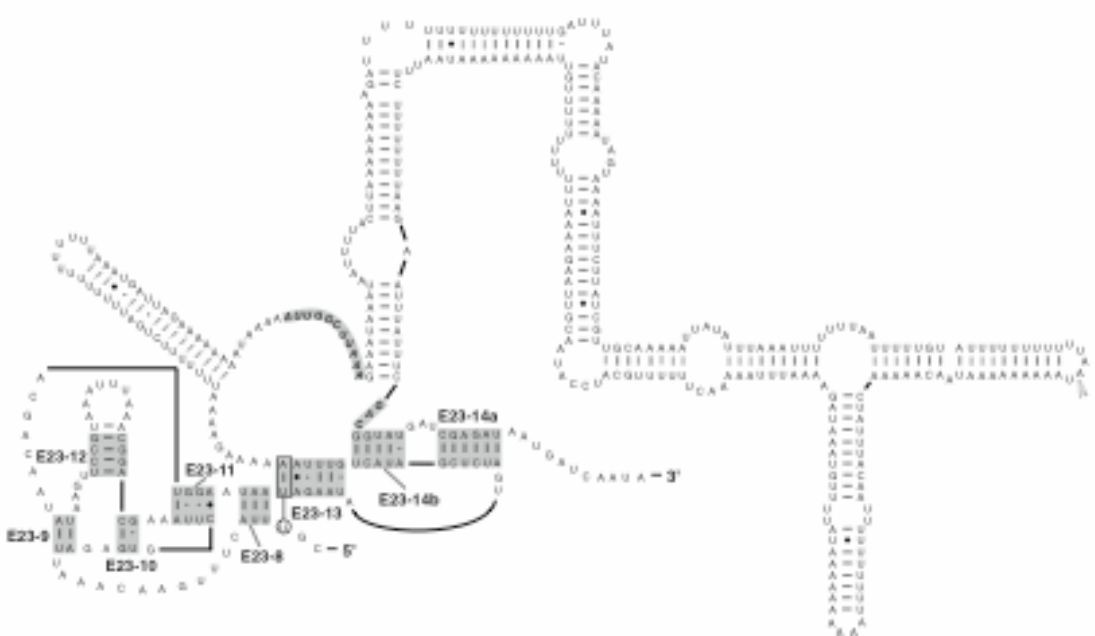
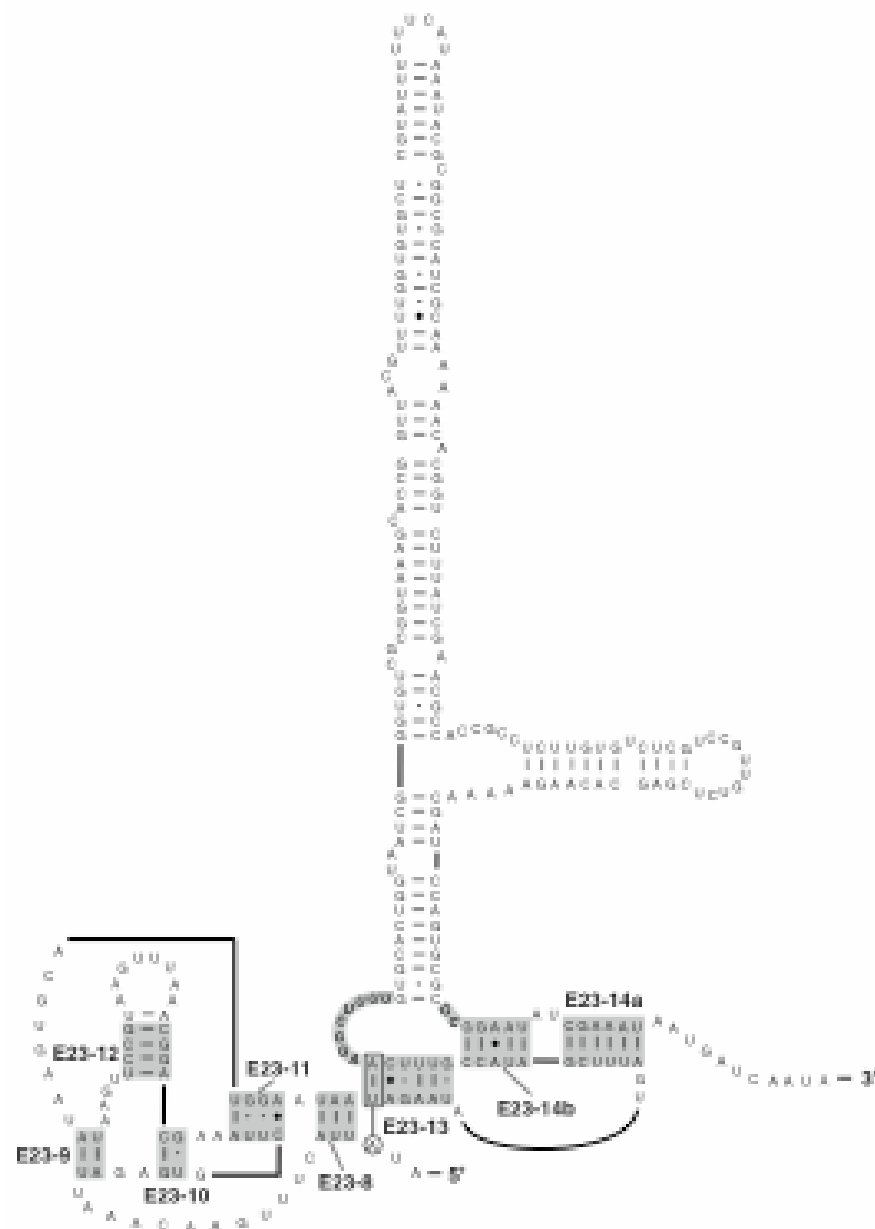


Figure 29 Continued.

I) *M. chobauti* X89441



J) *M. chobauti* AF423800

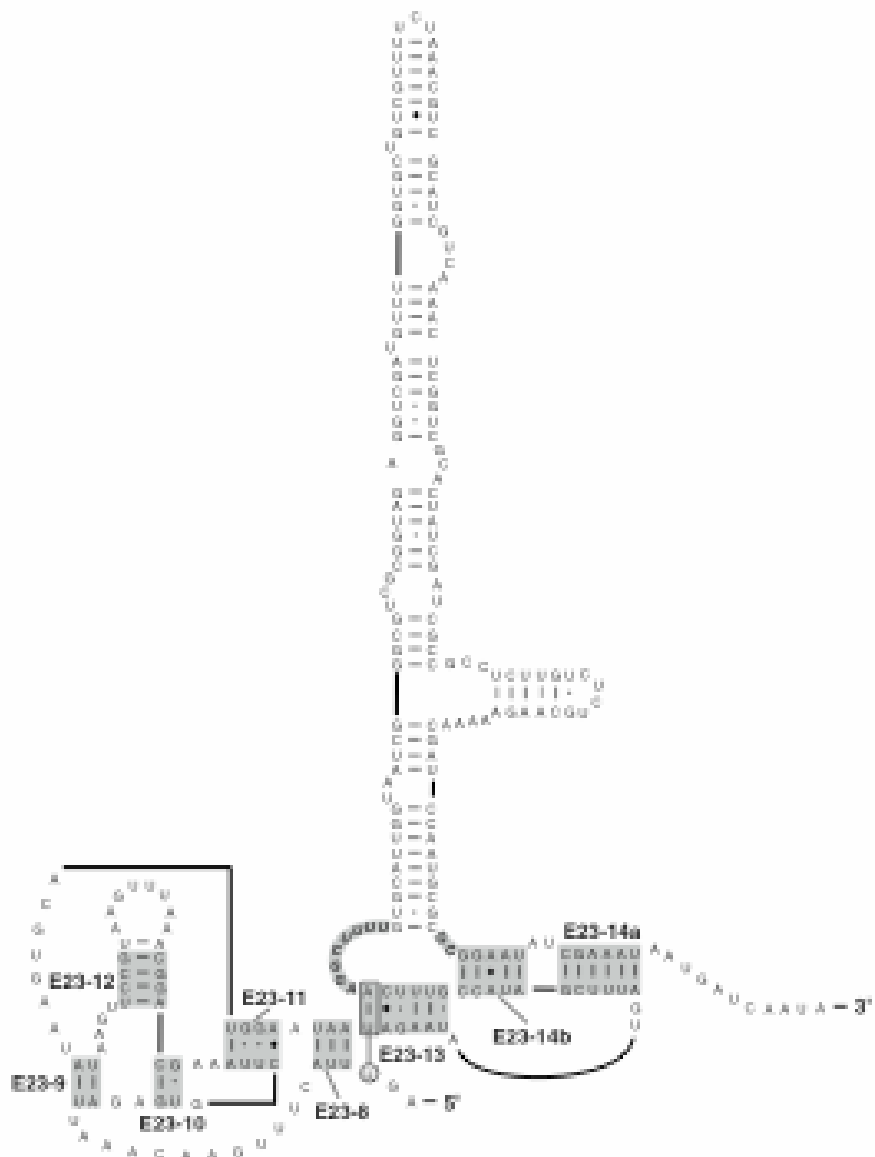


Figure 29 Continued.

Table 24. Distribution of sequence and structure within the strepsipteran insertion in the second hairpin loop of pseudoknot 2 within the variable region 4 (V4) of the 18S rRNA molecule.

Accession number	Taxon	5'-helix (?)/structure ^b	5'-conserved sequence ^{c,*}	3'-sequence/structure	3'-conserved sequence [*]
U65159	<i>Triozocera mexicana</i>	No.	UUUGAAAUUGGCUUAAA	238 nts; 87 bp	<u>GAC</u>
U65160	<i>Caenocholax fenyesi</i>	No.	UUUGAAAUUGGCUUAAA	235 nts; 87 bp	<u>GAC</u>
U65161	<i>Caenocholax fenyesi</i>	(2); 12, 18 bp.	UUUGAAAUUGGCUUAAA	43 nts; 17 bp	<u>GAC</u>
DQ026302 ^a	<i>Caenocholax fenyesi</i>	(2); 12, 18 bp.	UUUGAAAUUGGCUUAAA	43 nts; 17 bp	<u>GAC</u>
U65163	<i>Crawfordia</i> sp.	(1); 12 bp.	---AUUAUUGGCUUAAA	173 nts; 69 bp	<u>GAC</u>
X89440	<i>Stylops melittae</i>	(1); 22 bp.	--AGAAAUUGGCUUAAA	163 nts; 63 bp	<u>GAC</u>
X74763	<i>Xenos vesparum</i>	(1); 17 bp.	-AUAAA <u>UUGGCUUAAA</u>	284 nts; 109 bp	<u>GAC</u>
X77784	<i>Xenos vesparum</i>	(1); 17 bp.	-AUAAA <u>UUGGCUUAAA</u>	283 nts; 109 bp	<u>GAC</u>
U65164	<i>Xenos pecki</i>	(1); 12 bp.	--UAAA <u>UUGGCUUAAA</u>	329 nts; 118 bp	<u>GAC</u>
X89441	<i>Mengenilla chobauti</i>	No.	-----AGGCUUUU-	173 nts; 63 bp	<u>G-C</u>
AF423800	<i>Mengenilla chobauti</i>	No.	-----AGGCUUUU-	140 nts; 51 bp	<u>G-C</u>

^a Sequenced in this study

^b Number of bps within putative helices

^c Includes all unpaired nts flanking the 5'-end of conserved sequence.

^d Total bps in all putative helices

^{*} Bolded nts depict conserved sequences across Strepsiptera; underlined nts depict conserved sequences across all families but Mengenillidae.

sequences in the hairpin loop of pseudoknot 13/14 is summarized in Table 24. Because the highly conserved boundaries of pseudoknot 13/14 were not used as anchors by Choe *et al.* (1999b) for structure prediction, the structures they proposed are very different than our predictions, with the conservation of the abovementioned unpaired sequences involved in non-homologous basepairings (data not shown, Choe *et al.*, 1999b, their Fig. 3).

Given the lack of conservation in sequence and structure, it is likely that the insertion in pseudoknot 13/14 is part of the mature SSU rRNA and is probably not an intron. The major insertion points for introns in SSU rRNA have been well characterized (Wuyts *et al.*, 2001; Jackson *et al.*, 2002) and usually occur at the subunit interface or in conserved sites with known tRNA-rRNA interaction (Jackson *et al.*, 2002). Additionally, both the 18S rDNA and rRNA of *Xenos vesparum* were sequenced by Chalwatzis *et al.* (1995) and showed no differences in length. Given this, the functional significance of a peculiar insertion specific to Strepsiptera within a conserved pseudoknot across all Eukaryota remains unknown. However, recent evidence for a conserved sequence of the V4 forming a putative helix with a region in the V2 (our helix **H184b-1**) suggests a tertiary interaction between the two expansion segments is probable (Alkemar & Nygård, 2003), given their close proximity in the three-dimensional structure of the ribosome (Spahn *et al.*, 2001). Within our conserved sequence 5'-AUUGGCUUAAA-3', isolated by a variety of secondary structures (Fig. 29), the sequence 5'-AUUGGCUUA-3' can form a helix with the 5'-strand of helix **H184-1** and flanking nucleotides (Fig. 26). An analysis of basepair frequencies and

Table 25. Composition and degree of compensation for the base pairs of putative helices **H184-2** and **HV2-V4** of the 18S rRNA across 175 arthropods. For base composition percentages, bold values represent any base pair present at 2% or greater in the alignment. Asterisks denote positions that strictly covary for a given basepair (representing 3% or more of total basepair types), with the summed numbers providing a percentage of covariation. Note: percentages ignore phylogenetic correlation.

Helix ^a	bp ^b	# seq ^c	Base pair composition, % ^d																Gap ^e (-)
			Canonical						Non-canonical										
			GC	CG	UA	AU	GU	UG	AA	AC	AG	CA	CC	CU	GA	GG	UC	UU	
H184b																			
	1	176	----	0.6	89.7	----	----	----	----	----	----	----	----	----	----	----	----	----	9.7
	2	176	----	0.6	89.1	----	----	0.6	----	----	----	----	----	----	----	----	----	----	9.7
	3	176	----	0.6	----	86.3	----	----	----	0.6	2.9	----	----	----	----	----	----	----	9.7
	4	176	0.6	----	86.9	----	----	0.6	----	----	----	----	----	----	----	----	2.3	----	9.7
	5	176	----	0.6	69.1	----	----	19.4	----	----	----	----	0.6	----	----	----	----	0.6	9.7
	6	176	1.1	----	----	75.4	10.9	0.6	----	2.3	----	----	----	----	----	----	----	----	9.7
	7	176	----	77.1*	1.7	----	----	4.0*	----	0.6	----	6.3*	----	0.6	----	----	----	----	9.7
	8	176	----	87.4	----	----	----	----	----	----	1.1	0.6	----	----	----	0.6	----	0.6	9.7
	9	176	1.1	0.6	1.1	79.4	4.6	----	1.7	0.6	0.6	----	----	0.6	----	----	----	----	9.7
	10	176	0.6	0.6	1.1	71.4	0.6	----	4.0	1.7	5.1	----	----	----	----	1.1	----	0.6	11.4
HV2-V4																			
	1	176	1.7	----	----	78.9	0.6	----	2.3	4.6	----	----	----	----	----	----	----	----	0.6
	2	176	0.6	----	----	84.0	----	----	----	2.3	----	----	----	----	----	----	----	0.6	0.6
	3	176	1.1	----	----	79.4	6.3	----	----	----	----	----	----	----	0.6	----	----	----	0.6
	4	176	10.9*	----	----	68.6*	0.6	----	----	7.4*	0.6	----	----	----	----	----	----	0.6	0.6
	5	176	----	82.9	----	----	----	5.1	----	----	0.6	----	----	----	----	----	----	----	0.6
	6	176	----	86.9	----	----	----	0.6	----	----	1.1	----	----	----	----	----	----	----	0.6
	7	176	3.4*	----	1.1	81.1*	2.3	----	----	----	----	----	----	0.6	----	----	----	----	0.6
	8	176	0.6	----	1.1	82.3	1.1	----	----	----	----	----	0.6	----	----	----	----	0.6	0.6
	9	176	----	5.1*	3.4*	----	----	77.1*	----	----	1.1	----	----	----	0.6	----	----	----	0.6

^a Helix numbering refers the nucleotide positions shown in Figure 1.
^b Base pairs are numbered from 5'-end of 5'-strand of each helix.
^c Numbers vary at each position due to missing data (?), deletions (-) and possible presence of IUPAC-IUB ambiguity codes.
^d The first nucleotide is that in the 5'-strand.
^e Gaps represent single insertion or deletion events, not indels.
^f A covarying position is defined as having substitutions on both sides of the helix across the alignment.

degree of covariation for both helix **H184-1** and this putative tertiary helix, named here helix **HV2/V4**, reveals stronger support for the tertiary interaction (Table 25). However, because both structures can form across all Arthropoda, we cannot rule out the possibility that they both occur at different stages of ribosome assembly and function. It should be noted that in all non-strepsipteran arthropods, the formation of helix **HV2/V4** entails the dissolution of the first basepair in helix **E23-13**, thus providing evidence against the proposed base triple this basepair forms with the unpaired position immediately flanking the 5'-end of helix **E23-8** (Wuyts *et al.*, 2000).

E23-13/14 in published sequences

Previously, sequence heterogeneity was found in the 18S rRNA genes within single strepsipteran individuals (Whiting *et al.*, 1997). In that study, an automated alignment program (MALIGN, Wheeler & Gladstein, 1994) was used to align these strepsipteran sequences with 79 other sequences from the major lineages of insects for the purpose of estimating a phylogeny. Interestingly, the alignment of the majority of the insertion within hairpin 13/14 across these seven strepsipterans (Fig. 30) included two divergent sequences each from individuals of *C. fenyasi* Pierce 1909 and *Xenos pecki* (Whiting *et al.* 1997). The authors explained that there were sequencing problems for the V4 for these taxa due to the presence of multiple amplicons (Whiting *et al.*, 1997). Thus, the amplicons were cloned and sequenced, which resulted in two different sequences for both species. This result is in conflict with our analysis of seven strepsipteran taxa and our identification of two highly conserved short sequences within the variable secondary

structures in the V4 pseudoknot 13/14. Thus we conclude that this heterogeneity of the strepsipteran rDNA sequences is likely erroneous for the following reasons. First, the "long" sequence of *C. fenyesei* is probably *Triozocera mexicana* Pierce 1909 (Corioxenidae) (Fig. 30) because of high similarity in primary sequence (only one substitution and four indels out of 238 nts) and secondary structure (Table 24, Fig. 29A, C). Confusion of these sequences could have occurred through several means, nonetheless, our sequence of this region of the 18S rRNA from *C. f. texensis* (Fig. 26A) favors the probability that the "short" sequence of *C. fenyesei* (Fig. 29B) from Whiting *et al.* (1997) is correct.

Second, individuals of *X. pecki* do not contain "short" and "long" inserts in the loop of hairpin 13/14 because the "short" sequence appears to be an artifact of cloning. *X. pecki* "short" (Fig. 30) comprises three different sequences, of which only one should be aligned with the other strepsipteran sequences in the "insert 23" alignment of Whiting *et al.* (*Syst Biol* **46**: 56). The sequence in the dashed box (Fig. 30) depicts a region that is identical to *X. pecki* "long", except that this is part of the conserved alignment of Whiting *et al.* (*Syst Biol* **46**: 47). The regions that are boxed (Fig. 30) depict identical sequences that are misaligned due to the inclusion of the stretch of nucleotides from the unambiguous alignment misplaced in "insert 23". Finally, the remaining sequence (139 nts) of *X. pecki* "short", boxed and shaded dark (Fig. 30), is likely the 5'-end of the cloning vector. A sequence similarity search with the program BLAST (Altschul *et al.*, 1990) revealed four separate cloning vectors, all with 93% sequence identity (AF335420, AF335419, Y10545, U14118). Although a *X. pecki* "short" sequence

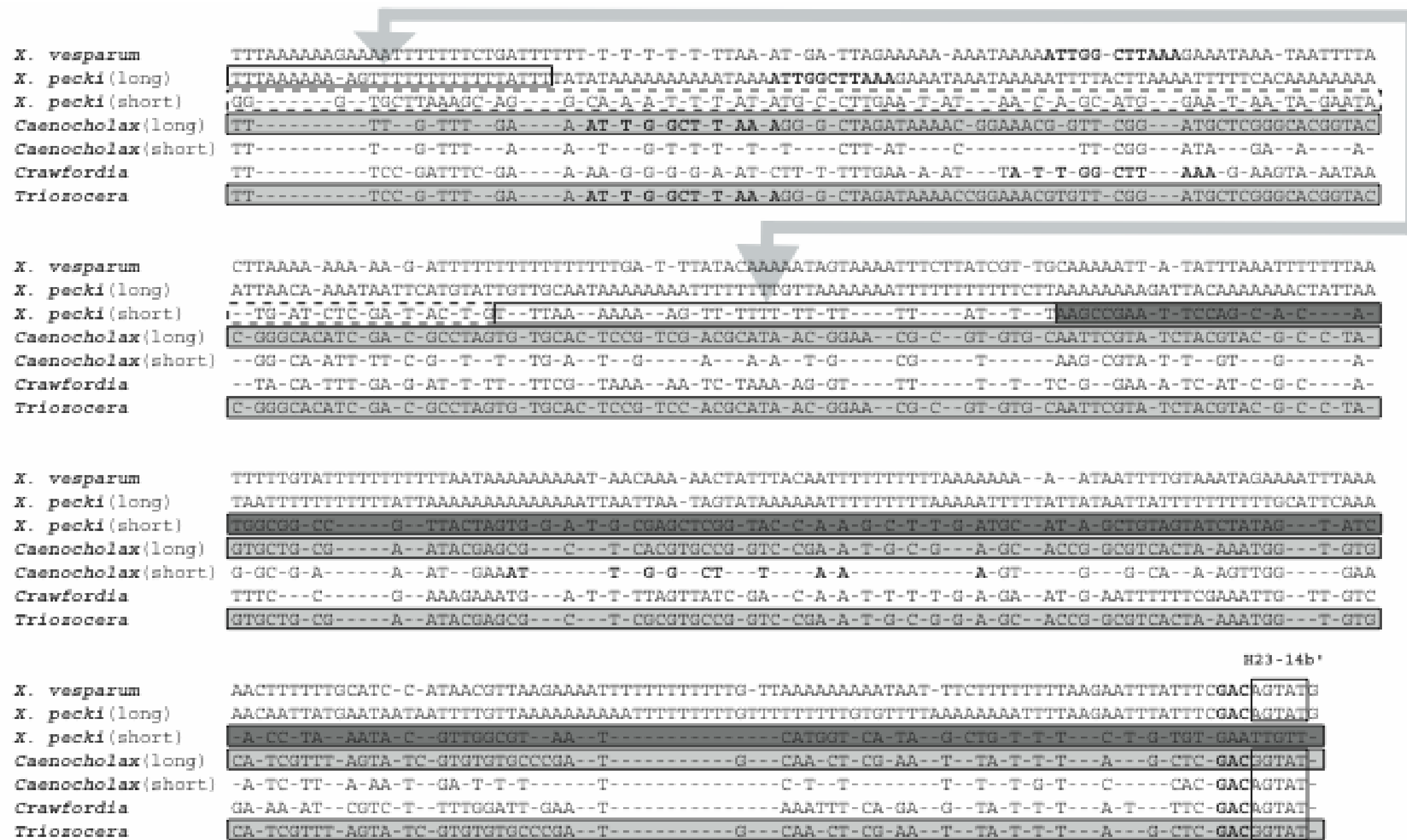


Figure 30. Recreated MALIGN alignment of Whiting *et al.* (1997, pg. 56). Sequences include the majority of the insertion depicted in Figure 3, plus the conserved 3'-strand of helix **E23-14b**. The conserved sequences described in the text are bolded. The dashed box in *X. pecki* (short) depicts a misaligned region of the 18S that should be in the unambiguously aligned data of Whiting *et al.* (1997, pg. 47). The boxes without shading show homologous (identical) sequences shared between *X. pecki* (long) and (short) as depicted with the grey arrow. The darkly shaded box depicts a sequence with 93% similarity to published cloning vectors. The lightly shaded boxes illustrate the near identity of *C. fenyasi* (long) with *T. mexicana*.

(minus the cloning vector) is still possible, verification awaits sequences of the 3'-strands of helices **E23-14a** and **E23-14b**, as well as the associated unpaired flanking regions (Fig. 29F).

The importance of using structure to align rRNA sequences

The elucidation of secondary structure of rRNA molecules guides the assignment of positional nucleotide alignment (Gutell *et al.*, 1985, 1992a, 1994; Kjer, 1995) and has been shown to improve phylogeny estimation (Dixon & Hillis, 1993; Kjer, 1995; Titus & Frost, 1996; Morrison & Ellis, 1997; Uchida *et al.*, 1998; Mugridge *et al.*, 1999; Cunningham *et al.*, 2000; Gonzalez & Labaree, 2000; Hwang & Kim, 2000; Lydeard *et al.*, 2000; Morin, 2000; Xia, 2000; Xia *et al.*, 2003; Kjer, 2004). The regions of length heterogeneous sequence alignments that are the most difficult to establish homology across often contain valuable phylogenetic signal (see Lee, 2001), and secondary structure provides an objective means for retrieving this information (see Gillespie, 2004). While often uninformative at the nucleotide level due to the rapidly evolving nature of rRNA variable regions when compared across highly divergent taxa, secondary structural characters can provide morphological evidence for similarity that is not immediately apparent with primary sequence data (reviewed in Gillespie *et al.*, 2004). It has been stated (e.g., Woese *et al.*, 1980; De Rijk *et al.*, 1994; Gutell & Damberger, 1996; Kjer, 1997) and demonstrated (Xia *et al.*, 2003; Kjer, 2004) that structural alignment also provides a means to "proofread" rRNA sequences for their accuracy, which is analogous to converting protein-encoding DNA sequences to their respective

amino acid sequences to check for shifts in the reading frame and unexpected stop codons. Our study also exemplifies the benefit of structure for the quality assessment of new and published data.

Implications for the systematic position of Strepsiptera

Given their peculiar biology and morphology, it is not surprising that the Strepsiptera are difficult to place phylogenetically within the Insecta. Some workers have allied the Strepsiptera with Coleoptera (beetles) based on the similarity of hind-wing-based flight (Kinzelbach, 1990; Kathirithamby, 1991b). However, several molecular phylogenetic estimations (all based on rDNA sequences) recover the strepsipterans as a sister taxon to the flies (Chalwatzis *et al.*, 1995, 1996; Whiting *et al.*, 1997). This is not surprising, since the exceedingly autapomorphic rDNA sequences of Strepsiptera are highly unlike those of other holometabolous insects (Gillespie, unpubl. data), and possibly group with dipteran rDNA sequences based on nucleotide convergence due to similar rapid rates of evolution (Hwang *et al.*, 1998). In fact, when convergence in nucleotide rates of substitution is accommodated for via character weighting or maximum likelihood modeling, the strepsipterans and dipterans are not recovered as a monophyletic group (Carmean & Crespi, 1995; Huelsenbeck, 1997; Hwang *et al.*, 1998). Interestingly, the majority of these phylogeny estimations have excluded the variable regions of insect rRNA molecules due to the difficulty in establishing alignment across heterogeneous sequences (but see Hwang *et al.*, 1998). The putative strepsipteran/dipteran synapomorphy of a doubly branched V7 region by Choe *et al.* (1999b) is not supported

in our study, as the V7 contains two helices (**H1118b** and **H1118c**; Fig. 26B) in many taxa within the Holometabola (Table 23).

Our predicted 18S rRNA structure model divides the V4, one of the most commonly sequenced molecules for arthropod phylogeny reconstruction (Cannone *et al.*, 2002), into 80 discrete regions defined by secondary structure (see Fig. 26 and alignment at <http://hymenoptera.tamu.edu/rna>). These partitions will be of use to a wide range of phylogenetic analysis programs which incorporate mixed evolutionary models. Additionally, we characterize the remaining regions of the 18S rRNA in concordance with published rRNA models and provide an alignment template that is custom for the Arthropoda. This should contribute greatly to future studies that employ structure into the process of homology assignment for the estimation of hexapod relationships.

The evolution of bizarre insertions in strepsipteran rDNA genes

The arthropod nucleotide insertions relative to other arthropod 18S rRNAs are usually confined to a few variable regions (usually V4 and V7; e.g., Crease & Taylor, 1998). However, strepsipterans have insertions in nearly every defined variable region of SSU rRNA, often with different base compositions in these regions when compared to other arthropod groups (Table 23). In addition, the strepsipterans have unique insertions in the highly conserved core rRNA structure (Choe *et al.*, 1999b). This suggests that the strepsipteran 18S rRNA genes are more tolerant of certain mutations rather than selected against or pruned by processes of gene conversion and/or unequal crossing over (Arnheim *et al.*, 1980; Ohta, 1980; Dover, 1982; Arnheim, 1983; Flavell, 1986;

Nagylaki, 1988). High rates of nucleotide substitution occur in strepsipterans (Hwang *et al.*, 1998), and perhaps the rate of mutation exceeds the rate at which gene conversion and/or unequal crossing over can remove novel insertion events, leading to their eventual fixation (Ohta, 1982). The fixation of insertion events relative to their removal by purifying selection would be directly affected by the size of the strepsipteran genome and the number of copies of rDNA genes.

In organisms with rapid rates of nucleotide evolution, a low rDNA copy number would allow gene conversion and/or unequal crossing over to keep the highly evolving rDNA copies concerted. Unfortunately, the rDNA copy number for any strepsipteran genome has yet to be established. However, a recent study determined that the genome size of *C. f. texensis* has one of the smallest C-values of any animal (Johnston *et al.*, 2004). Since rDNA copy number is usually positively correlated with genome size (Prokopowich *et al.*, 2003), it is likely that *C. f. texensis* has a low rDNA copy number. Further support for this comes from the evidence that *C. f. texensis* undergoes endoreduplication, and can have up to 16 copies of its genome in male flight muscle and in the tissues that support the hundreds of thousands of rapidly developing embryos in the mature female (Johnston *et al.*, 2004). It has been shown that, despite having only one rDNA gene per genome, the ciliated protozoan *Tetrahymena thermophila* endoreduplicates its genome approximately 200 fold (Gall, 1974; Yao *et al.*, 1974). Thus, the ability to endoreduplicate certain regions of the genome would allow for organisms with low copies of rDNA arrays to still produce enough rRNAs to meet the demands of the cell. In the case of strepsipterans, low rDNA copy number and

endoreduplication could possibly combat the rapid rates in nucleotide evolution by accommodating the homogenization process of the rDNA copies.

We present this hypothesis for a low rDNA copy number in strepsipterans in light of recent controversy regarding the evolution of genome size. It has been proposed that variation in genome size can be indicative of biases in small insertions and deletions (indels) (Petrov *et al.*, 1996; Petrov *et al.*, 2000; Bensasson *et al.*, 2001; Petrov, 2001; Petrov, 2002). In general, these studies ascribe to the logic that larger genomes tend to accumulate more insertions over time, with smaller genomes likely riddled with more deletion events. However, these generalizations of genome size are probably more complicated (Gregory, 2003, 2004), and there are likely other selective factors, such as cell volume (Bennett, 1972; Cavalier-Smith, 1985; Gregory, 2001) and endoreduplication (Nagl, 1978), responsible for shaping genome size. The strepsipterans certainly pose a paradox to the C-value enigma (Gregory, 2003), having one of the smallest animal genomes and possessing some of the largest insertions in all of the documented SSU rRNA sequences. Further study is needed to understand the tolerance strepsipterans have for accumulating large insertions within an otherwise tiny genome.

Experimental procedures

Taxa examined

The majority of the published 18S rRNA sequences used in this study were compiled from two recent phylogenetic studies on insects (Kjer, 2004) and arthropods (Mallatt *et al.* 2004). A taxon list with respective Genbank accession numbers and other

information can be found at the jRNA website. The accession number for *C. f. texensis* is DQ026302. The voucher specimens were deposited in either the Texas A&M insect collection or the Museum of Natural History, Oxford.

Genome isolation, PCR, and sequencing

For the sequence generated in this study, total genomic DNA was isolated using DNeasy™ Tissue Kits (Qiagen). PCR conditions followed those of Cognato & Vogler (2001). A complete list of previously published primers, as well as newly designed primers specific to *C. f. texensis*, is posted at the jRNA website. Double-stranded DNA amplification products were sequenced directly with ABI PRISM™ (Perkin-Elmer) Big Dye Terminator Cycle Sequencing Kits and analyzed on an Applied Biosystems (Perkin-Elmer) 377 automated DNA sequencer. Both anti-sense and sense strands were sequenced for all taxa, and edited manually with the aid of Sequence Navigator™ (Applied Biosystems). During editing of each strand, nucleotides that were readable, but showed either irregular spacing between peaks, or had some significant competing background peak, were coded with lower case letters or IUPAC-IUB ambiguity codes. Consensus sequences were exported into Microsoft Word™ for manual alignment.

Multiple sequence alignment

Our *C. f. texensis* 18S sequence, as well as 11 other published strepsipteran sequences and the panarthropod taxa from Mallatt *et al.* (2004), were aligned to the recent structural alignment of Kjer (2004). Adjustments to Kjer's alignment were made either

in strict adherence to the 16S-like models on the Comparative RNA Website (Gutell *et al.* 1994; Cannone *et al.*, 2002) or from information provided by covariation analysis (see below). Additionally, the V4 region was realigned according to the model of Wuyts *et al.* (2000), with the tertiary interaction between the V2 and V4 included (Alkemar & Nyågrd, 2003). The structural notation of the alignment followed Gillespie *et al.* (2004). Length variable sequences, especially hairpin-stem loops, were evaluated in the program *mfold* (version 3.1; <http://bioinfo.math.rpi.edu/~zukerm/>), which folds rRNA based on free energy minimizations (Mathews *et al.*, 1999; Zuker *et al.*, 1999). These free energy-based predictions were used to facilitate the search for potential base-pairing stems, which were confirmed only by the presence of compensatory base changes across a majority of taxa. Thermodynamic-based folding algorithms usually predict several plausible sub-optimal structure models in addition to the optimal one, and in many situations the difference in energy value between the optimal and sub-optimal is very small. Thus, all of these predicted structures should be considered, not just the optimal one. Consequently, we are more confident of a predicted structure that contains certain sequence/structure motifs characteristic of rRNA, such as hairpin loops with three U nucleotides, YUCG and GNRA tetra-hairpin loops (Woese *et al.*, 1990b) and AA and AG juxtapositions at the ends of helices (Elgavish *et al.*, 2001). Regions of the alignment wherein homology assignments could not be made with a high level of confidence were treated following the methodology of Gillespie (2004).

Alignment-based statistics and structure diagrams

Our alignment was modified into a Nexus file for further manipulation using scripts available at the jRNA website. Scripts were used to calculate basepair-frequency tables (providing a percentage of covariation for each basepair within a putative helix), nucleotide composition in non-pairing regions of the alignment, and mean length ranges for the variable regions throughout the 18S rRNA. Secondary structure diagrams were generated with the computer program XRNA (developed by B. Weiser and H. Noller, University of Santa Cruz) and adjusted manually for production of the figures. All alignment formats, alignment-based statistics, structure diagrams (including a panarthropod consensus model), and scripts used to parse and analyze the data are available at the jRNA website following the links to 'arthropoda'.

CHAPTER VII

SUMMARY

In this dissertation, I demonstrated that higher order structure can be predicted from multiple sequence alignments to: 1., improve homology assignment and provide an objective criterion for data exclusion (a "conditional combination" approach for phylogeny estimation), 2. provide information about the sequenced molecules that allow for sub-partitions of the datasets to be create and modeled as independent character classes ("stems and loops"), 3. improve the existing knowledge of the structure and function of the rRNA and ultimately the ribosome, while often identifying novel structural features, and 4. identify sequencing artifacts on public genetic databases that were previously undetected without structural inference. My dissertation will be useful for evolutionary biologists concerned with the structure, function, and evolution of rRNA, as well as systematists interested in structure-based applications for the phylogenetic analysis of these intriguing molecules. Much of the philosophy and methodology behind the experiments conducted in this dissertation are explained at the jRNA website (<http://hymenoptera.tamu.edu>), an on-going project established in collaboration with Matt Yoder (TAMU). Matt has created invaluable tools, such as Perl scripts for parsing alignments for various informatics platforms following my structural convention, and alignments summary statistics, that have already proven to facilitate all aspects of rRNA structure-based studies. For more information on the studies within

this dissertation, as well as many other studies on rRNA, the reader is encouraged to continually check the jRNA for new developments (Svedberg).

REFERENCES

- Alkemar, G. and Nygård, O. (2003). A possible tertiary rRNA interaction between expansion segments ES3 and ES6 in eukaryotic 40S ribosomal subunits. *RNA* **9**: 20-24.
- Allard, E. (1860) Essai monographique sur les Galerucites Anisopodes (Latr.) ou description des Altises d'Europe et des bords de la mer Méditerranée. *Annales de la Société Entomologique de France* **8**: 39-144.
- Allard, E. (1866) Monographie des Alticides tribu de la famille des Phytophages. *L'Abeille, Journal d'Entomologie* **3** **53**: 169-417.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Alvares, L.E., Wuyts, J., Van de Peer, Y., Silva, E.P., Coutinho, L.L., Brison, O. and Ruiz, I.G.R. (2004) The 18S rRNA from *Odontophrynus americanus* 2n and 4n (Amphibia: Anura) reveals unusual extra sequences in the variable region V2. *Genome* **47**: 421-428.
- Amako, D., Kwon, O.-Y. and Ishikawa, H. (1996) Nucleotide sequence and presumed secondary structure of the 28S rRNA of pea aphid: Implication for diversification of insect rRNA. *J Mol Evol* **43**: 469-475.
- Arnheim, N. (1983) Concerted evolution of multigene families. In *Evolution of Genes and Proteins* (Nei, M. and Koehn, R.K., eds). pp. 38-61. Sinauer, Sunderland, MA.
- Arnheim, N., Krystal, M., Shmickel, R., Wilson, G., Ryder, O. and Zimmer, E. (1980) Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proc Natl Acad Sci USA* **77**: 7323-7327.
- Askew, R.R. and Shaw, M.R. (1986) Parasitoid communities: Their size, structure and development. In *Insect Parasitoids* (Waage, J. and Greathead, D., eds), 13th Symposium of the Royal Entomological Society of London, 18-19 September 1985, pp. 225-264.
- Baldwin, B.G., Sanderson, M.J., Porter, J.M., Wojciechowski, M.F., Campbell, C.C. and Donoghue, M.J. (1995) The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Ann Mo Bot Gard* **82**: 257-277.

- Bechyné, J. and Springlova de Bechyné, B. (1976) Notes sur quelques Aphthonini nouveaux ou peu connus. (Chrysomeloidea - Alticinae) Coleoptera, Phytophaga. *Pesquisas* **28**: 1-15.
- Belshaw, R. and Quicke, D.L.J. (2002) Robustness of ancestral state estimates: Evolution of life history strategy in ichneumonoid parasitoids. *Syst Biol* **51**: 450-477.
- Bennett, M.D. (1972) Nuclear DNA content and minimum generation time in herbaceous plants. *Proc Roy Soc Lond B Biol Sci* **181**: 109-135.
- Bensasson, D., Petrov, D.A., Zhang, D.-X., Hartl, D.L. and Hewitt, G.M. (2001) Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Mol Biol Evol* **18**: 246-253.
- Billboud, B., Geurrucci, M.-A., Masselot, M. and Misof, B. (2000) Cirripede phylogeny using a novel approach: Molecular morphometrics. *Mol Biol Evol* **17**: 1435-1445.
- Blake, D.H. (1958) A review of some galerucine beetles with excised middle tibia in the male. *Proc US Natl Mus* **108**: 59-106.
- Böving, A.G. and Craighead, F.C. (1931) An illustrated synopsis of the principal larval forms of the order Coleoptera. *Entomol Amer (N.S.)* **11**: 1-351.
- Bremer, K. (1988) The limits of amino acids sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**: 795-803.
- Campbell, B.C., Steffen-Campbell, J.D. and Gill, R.J. (1994) Evolutionary origin of whiteflies (Hemiptera: Sternorrhyncha: Aleyrodidae) inferred from 18S rDNA sequences. *Insect Mol Biol* **3**: 175-194.
- Campbell, B.C., Steffen-Campbell, J.D., Sorensen, J.T. and Gill, R.J. (1995) Paraphyly of Homoptera and Auchenorrhyncha inferred from 18S rDNA nucleotide sequences. *Syst Entomol* **20**: 175-194.
- Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pande, N., Shang, Z., Yu, N. and Gutell, R.R. (2002) The Comparative RNA Web (CRW) Site: An online database of comparative sequence and structure information for ribosomal, intron and other RNAs. *BMC Bioinformatics* **3**:2. [Correction: *BMC Bioinformatics* **3**:15.]
- Carmean, D. and Crespi, J.B. (1995) Do long branches attract flies? *Nature* **373**: 666.

- Cavalier-Smith, T. (1985) Cell volume and the evolution of eukaryotic genome size. In *The Evolution of Genome Size* (Cavalier-Smith, T., ed), pp. 104-184. John Wiley & Sons, Chichester, England.
- Chalwatzis, N., Baur, A., Stetzer, E., Kinzelbach, R. and Zimmermann, R.K. (1995) Strongly expanded 18S ribosomal-RNA genes correlated with a peculiar morphology in the insect order of Strepsiptera. *Zool Anal Complex Systems* **98**: 115-126.
- Chalwatzis, N., Hauf, J., Van de Peer, Y., Kinzelbach, R. and Zimmermann, R.K. (1996) 18S ribosomal-RNA genes of insects: Primary structure of the genes and molecular phylogeny of the Holometabola. *Ann Entomol Soc Am* **89**: 788-803.
- Chapuis, M.F. (1875) Famille des Phytophages, Vol. II. In *Histoire Naturelle des Insectes. Genera des Coléoptères ou exposé méthodique et critique de tous les genres proposés jusqu'ici dans cet ordre d'Insectes Tome onzieme* (Lacordaire, T., ed.), 420 pp., pls. 124-174. A la Librairie Encyclopédique de Roret, Paris, France.
- Choe, C.P., Hancock, J.M., Hwang, U.W. and Kim, W. (1999a) Analysis of the primary sequence and secondary structure of the unusually long SSU rRNA of the soil bug, *Armadillidium vulgare*. *J Mol Evol* **49**: 798-805.
- Choe, C.P., Hwang, U.W. and Kim, W. (1999b) Putative secondary structures of unusually long strepsipteran SSU rRNAs and its phylogenetics implications. *Mol Cells* **9**: 191-199.
- Clark, C.G. (1987) On the evolution of ribosomal RNA. *J Mol Evol* **25**: 343-350.
- Clark C.G., Tague, B.W., Ware, V.C. and Gerbi, S.A. (1984) *Xenopus laevis* 28S ribosomal RNA: A secondary structural model and its evolutionary and functional implications. *Nucleic Acids Res* **12**: 6197-6220.
- Cognato, A.I. and Vogler, A.P. (2001) Exploring data interaction and nucleotide alignment in a multiple gene analysis of *Ips* (Scolytinae). *Syst Biol* **50**: 758-780.
- Collins, L.J., Moulton, V. and Penny, D. (2000) Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J Mol Evol* **51**: 194-204.
- Cox, D.R. (1962) Further results on tests of families of alternate hypotheses. *J R Stat Soc B* **24**: 406-424.
- Crandall, K.A. and Fitzpatrick, J.J. (1996) Crayfish molecular systematics: Using a combination of procedures to estimate phylogeny. *Syst Biol* **45**: 1-26.

- Crease, T.J. and Colbourne, J.K. (1998) The unusually long small-subunit ribosomal RNA of the crustacean, *Daphnia pulex*: Sequence and predicted secondary structure. *J Mol Evol* **46**: 307-313.
- Crease, T.J. and Taylor, D.J. (1998) The origin and evolution of variable-region helices in V4 and V7 of the small-subunit ribosomal RNA of branchiopod crustaceans. *Mol Biol Evol* **15**: 1430- 1446.
- Crowson, R.A. (1955) *The Natural Classification of the Families of Coleoptera*. Nathaniel Lloyd & Co. London. 187 pp.
- Crowson, R.A. and Crowson, E.A. (1996) The phylogenetic relations of Galerucinae-Alticinae. In *Chrysomelidae Biology vol. 1: The Classification, Phylogeny and Genetics* (Jolivet, P.H. and Cox, M.L., eds), pp. 97-118. SPB Publishing, Amsterdam, the Netherlands.
- Cunningham, C.O., Aliesky, H. and Collins, C.M. (2000) Sequence and secondary structure variation in the *Gyrodactylus* (Platyhelminthes: Monogenea) ribosomal RNA gene array. *J Parasitol* **86**: 567-576.
- Dahlberg, A.E. (1989) The functional role of ribosomal RNA in protein synthesis. *Cell* **57**:525-529.
- De Rijk, P., Van de Peer, Y., Chapelle, S. and De Wachter, R. (1994) Database on the structure of large ribosomal subunit RNA. *Nucleic Acids Res* **22**: 3495-3501.
- DeRijk, P., Van de Peer, Y., Van den Broeck, I. and De Wachter, R. (1995) Evolution according to large ribosomal subunit RNA. *J Mol Evol* **41**: 366-375.
- Dixon, M.T. and D.M. Hillis. (1993) Ribosomal secondary structure: Compensatory mutations and implications for phylogenetic analysis. *Mol Biol Evol* **10**: 256-267.
- Doguet, S. (1994) *Coléoptères Chrysomelidae, Volume 2, Alticinae*. Federation Francaise des Societes de Sciences Naturelles. Paris. 694 pp.
- Donaghue, M.J., Olmstead, R.G., Smith, J.F. and Palmer, J.D. (1992) Phylogenetic relationships of Dipsacales based on *rbcl* sequences. *Ann Miss Bot Gard* **79**: 333-345.
- Douzery, E. and Catzeflis, F.M. (1995) Molecular evolution of the mitochondrial 12S rRNA in Ungulata (Mammalia). *J Mol Evol* **41**: 622-636.
- Dover, G.A. (1982) Molecular drive: A cohesive mode of species evolution. *Nature* **299**: 111-117.

- Duckett, C.N., Gillespie, J.J. and Kjer, K.M. (2004) Relationships among the subfamilies of Chrysomelidae inferred from small subunit ribosomal DNA, with special emphasis on the relationship between the flea beetles and the Galerucinae. In *New Developments in the Biology of Chrysomelidae* (Jolivet, P.H., Santiago-Blay, J.A. and Schmitt, M., eds), pp. 3-18. Kluwer Academic Publ., Boston.
- Elgavish, T., Cannone, J.J., Lee, J.C., Harvey S.C. and Gutell R.R. (2001) AA.AG@Helix.Ends: A:A and A:G Base-pairs at the Ends of 16S and 23S rRNA Helices. *J Mol Biol* **310**: 735-753.
- Farrell, B.D. (1998) "Inordinate fondness" explained: why are there so many beetles? *Science*, **281**, 555-559.
- Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783-791.
- Flavell, R.B. (1986) Structure and control of expression of ribosomal RNA genes. *Oxf Surv Plant Mol Cell Biol* **3**: 252-274.
- Flores-Villela, O., Kjer, K.M., Benabib, M. and Sites, J.W. (2000) Multiple data sets, congruence and hypothesis testing for the phylogeny of basal groups of the lizard genus *Sceloporus* (Squamata, Phrynosomatidae). *Syst Biol* **49**: 713-739.
- Fontana, W., Konings, D.A.M., Stadler, P.F. and Schuster, P. (1993) Statistics of RNA secondary structures. *Biopolymers* **33**: 1389-1404.
- Fresco, J.R., Alberts, B.M. and Doty, P. (1960) Some molecular details of the secondary structure of ribonucleic acid. *Nature* **188**: 98.
- Friedrich, M. and Tautz, D. (1997a) An episodic change of rDNA nucleotide substitution rate has occurred at the time of the emergence of the insect order Diptera. *Mol Biol Evol* **14**: 644-653.
- Friedrich, M. and Tautz, D. (1997b) Evolution and phylogeny of the Diptera: A molecular phylogenetic analysis using 28S rDNA sequences. *Syst Biol* **46**: 674-698.
- Furth, D.G. and Suzuki, K. (1994) Character correlation studies of problematic genera of Alticinae in relation to Galerucinae (Coleoptera: Chrysomelidae). In *Proceedings of the Third International Symposium on the Chrysomelidae, Beijing* (Furth, D.G. ed.), pp. 116-135. Backhuys Publishers, Leiden, the Netherlands.
- Gall, J.G. (1974) Free ribosomal RNA genes in the macronucleus of *Tetrahymena*. *Proc Nat Acad Sci USA* **71**: 3078-3081.

- Gatesy, J., Hayashi, C., DeSalle, R. and Vrba, E. (1994) Rate limits for pairing and compensatory change: The mitochondrial ribosomal DNA of antelopes. *Evolution* **48**: 188-196.
- Gautheret, D., Damberger, S.H. and Gutell, R.R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J Mol Biol* **248**: 27-43.
- Gerbi, S.A. (1985) Evolution of ribosomal DNA. In *Molecular Evolutionary Genetics* (MacIntyre, R.J., ed), pp. 419-517. Plenum, New York.
- Gibson, A., Gowri-Shankar, Higgs, P.G. and Rattray, M. (2005) A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. *Mol Biol Evol* **22**: 251-264.
- Gillespie, J.J. (2001) Inferring phylogenetic relationships among basal taxa of the leaf beetle tribe Luperini (Chrysomelidae: Galerucinae) through the analysis of mitochondrial and nuclear DNA sequences. Unpublished M.S. Thesis. University of Delaware, Newark.
- Gillespie, J.J. (2004) Characterizing regions of ambiguous alignment caused by the expansion and contraction of hairpin-stem loops in ribosomal RNA molecules. *Mol Phylogenet Evol* **33**: 936-943.
- Gillespie, J.J., Duckett, C.N. and Kjer, K.M. (2001) Identification of a gene region that gives good phylogenetic signal for determining high level divergences within alticine and galerucine chrysomelids. *Chrysomela*, **40/41**, 10-11. Available at http://www.coleopsoc.org/chrys/chrysomela_4041r.pdf
- Gillespie, J.J., Kjer, K.M., Duckett, C.N. and Tallamy, D.W. (2003) Convergent evolution of cucurbitacin feeding in spatially isolated rootworm taxa (Coleoptera: Chrysomelidae; Galerucinae, Luperini). *Mol Phylogenet Evol* **29**: 161-175.
- Gillespie, J.J., Kjer, K.M., Riley, E.R. & Tallamy, D.W. (2004a) The evolution of cucurbitacin pharmacophagy in rootworms: insight from Luperini paraphyly. In *New Developments in the Biology of Chrysomelidae* (Jolivet, P.H., Santiago-Blay, J.A. and Schmitt, M., eds), pp. 37-57. Kluwer Academic Publ., Boston.
- Gillespie, J.J., Cannone, J.J., Gutell, R.R. and Cognato, A.I. (2004b) A secondary structural model of the 28S rRNA expansion segments D2 and D3 from rootworms and related leaf beetles (Coleoptera: Chrysomelidae; Galerucinae). *Insect Mol Biol* **13**: 495-518.

- Gillespie, J.J., Yoder, M.J. and Wharton, R.A. (2005a) Predicted secondary structures for 28S and 18S rRNA from Ichneumonoidea (Insecta: Hymenoptera: Apocrita): Impact on sequence alignment and phylogeny estimation. *J Mol Evol* In press.
- Gillespie, J.J., Munro, J.B., Heraty, J.M., Yoder, M.J., Owen, A.K., Carmichael, A.E. (2005b) A secondary structural model of the 28S rRNA expansion segments D2 and D3 for chalcidoid wasps (Hymenoptera: Chalcidoidea). *Mol Biol Evol* **22**: 1593-1608.
- Giribet, G. and Wheeler, W.C. (2001) Some unusual small-subunit ribosomal RNA sequences of metazoans. *Amer Mus Novitates* **3337**: 1-14.
- Goertzen, L.R., Cannone, J.J., Gutell, R.R. and Jansen, R.K. (2003) ITS secondary structure derived from comparative analysis: Implications for sequence alignment and phylogeny of the Asteraceae. *Mol Phylogenet Evol* **29**: 216-234.
- Goldman, N. (1993) Statistical tests of models of DNA substitution. *J Mol Evol* **36**: 182-198.
- Goloboff, P.A. (1999) Analyzing large data sets in reasonable times: Solutions for composite optima. *Cladistics* **15**: 415-428.
- Goloboff, P.A., Farris, J. and Nixon, K. (2003) T.N.T.: Tree Analysis Using New Technology. Program and documentation, available from the authors and at www.zmuk.dk/public/phylogeny.
- Gonzalez, P. and Labarere, J. (2000) Phylogenetic relationships of *Pleurotus* species according to the sequence and secondary structure of the mitochondrial small-subunit rRNA V4, V6 and V9 domains. *Microbiology* **146**: 209-221.
- Gorski, J.L., Gonzalez, K.L. and Schmicuez, R.D. (1987) The secondary structure of human 28S rRNA: The structure and evolution of a mosaic rRNA gene. *J Mol Evol* **24**: 236-251.
- Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cell size and the C-value enigma. *Biol Rev* **76**: 65-101.
- Gregory, T.R. (2003) Is small indel bias a determinant of genome size? *Trends Genet* **19**: 485-488.
- Gregory, T.R. (2004) Insertion-deletion bias and the evolution of genome size. *Gene* **324**: 15-34.

- Gressit, J.L. and Kimoto, S. (1963) The Chrysomelidae (Coleoptera) of China and Korea. Part 2. *Pacific Insects Monograph* **1b**: 743-893.
- Gruev, B. and Tomov, V. (1986) Coleoptera, Chrysomelidae. Part II. Chrysomelinae, Galerucinae, Alticinae, Hispinae, Cassidinae. *Fauna of Bulgaria* **16**: Sofia, Bulgaria. 388 pp.
- Gutell, R.R. (1992) Evolutionary characteristics of 16S and 23S rRNA structures. In *The Origin and Evolution of Prokaryotic and Eukaryotic Cells* (Hartman, H. and Matsuno, K. eds), pp. 243-309. World Scientific Publishing Co., Hackensack, NJ.
- Gutell, R.R. (1993) Collection of Small Subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res* **21**: 3051-3054.
- Gutell, R.R. (1994) Collection of Small Subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucleic Acids Res* **22**: 3502-3507.
- Gutell, R.R. (1996) Comparative sequence analysis and the structure of 16S and 23S rRNA. In *Ribosomal RNA Structure, Evolution, Processing and Function in Protein Synthesis* (Dahlberg, A.E. and Zimmerman, E.A. eds), pp. 111-129. CRC Press, Boca Raton, FL.
- Gutell, R.R., Weiser, B., Woese, C.R. and Noller, H.F. (1985) Comparative anatomy of 16S-like ribosomal RNA. *Prog Nucleic Acid Res Mol Biol* **32**: 155-216.
- Gutell, R.R. and Fox, G.E. (1988) A compilation of large subunit RNA sequences presented in a structural format. *Nucleic Acids Res* **16S**: r175-r269.
- Gutell, R.R., Schnare, M.N. and Gray, M.W. (1990) A compilation of large subunit (23S-like) ribosomal RNA sequences presented in a secondary structure format. *Nucleic Acids Res* **18S**: 2319-2330.
- Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J. and Stormo, G.D. (1992a) Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res* **20**: 5785-5795.
- Gutell, R.R., Schnare, M.N. and Gray, M.W. (1992b) A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucleic Acids Res* **20S**: 2095-2109.
- Gutell, R.R., Gray, M.W. and Schnare, M.N. (1993) A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures: 1993. *Nucleic Acids Res* **21S**: 3055-3074.

- Gutell, R.R., Larsen, N. and Woese, C.R. (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Rev* **58**: 10-26.
- Gutell, R.R. and Damberger, S.H. (1996) Comparative sequence analysis of experiments performed during evolution. In *Ribosomal RNA and Group I Introns* (Green, R. and Schroeder, R., eds), pp. 15-33. R.G. Landes Company, Austin, TX.
- Gutell, R.R., Lee, J.C. and Cannone, J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol* **12**: 301-310.
- Hancock, J.M. and Dover, G.A. (1988) Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. *Mol Biol Evol* **5**: 377-392.
- Hancock, J.M., Tautz, D. and Dover, G.A. (1988) Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. *Mol Biol Evol* **5**: 393-414.
- Hancock, J.M. and Vogler, A.P. (2000) How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: Implications for phylogeny reconstruction. *Mol Phylogenet Evol* **14**: 366-374.
- Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **42**: 160-174.
- Hassouna, N., Michot, B. and Bachellerie, J.-P. (1984) The complete nucleotide sequence of mouse 28S rRNA gene: Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Res* **12**: 3563-3583.
- Hastings, W. (1970) Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57**: 97-109.
- Heikertinger, F. (1912) Unterfamilie: Halticinae. In: *Fauna Germanica, Die Käfer des Deutschen Reiches*, Vol. 4 (Reitter, E., ed), pp. 143-212. W. Junk, Stuttgart, Germany.
- Heikertinger, F. (1924) Die Halticinengenera der Palaearktis und Nearktis. Bestimmungstabelle. *Koleopterologische Rundschau* **9**: 25-70.
- Heikertinger, F. (1941) Bestimmungsbellen europäischer Käfer. LXXXII. Fam. Chrysomelidae. 5. Subfam. Halticinae. Bestimmungstabelle der gattungen der Palaerktischen Halticinen. *Ibidem* **26**: 67-89.

- Heikertinger, F. and Csiki, E. (1939) Chrysomelidae: 11. Halticinae I, Vol. 25 (Pars 166). In *Coleopterum Catalogus 166* (Junk, W., Schenkling, S., eds.), pp. 1-336. W. Junk, s'Gravenhage, the Netherlands.
- Heikertinger, F. and Csiki, E. (1940) Chrysomelidae: 11. Halticinae II, Vol. 25 (Pars 166). In *Coleopterum Catalogus 166* (Junk, W., Schenkling, S., eds.), pp. 337-635. W. Junk, s'Gravenhage, the Netherlands.
- Hibbet, D.S., Fukumasa-Nakai, Y., Tsuneda, A. and Donoghue, M.J. (1995) Phylogenetic diversity in shiitake inferred from nuclear ribosomal DNA. *Mycologia* **87**: 618-638.
- Hickson, R.E., Simon, C., Cooper, A., Spicer, G.S., Sullivan, J. and Penny, D. (1996) Conserved sequence motifs, alignment and secondary structure for the third domain of animal 12S rRNA. *Mol Biol Evol* **13**: 150-169.
- Hickson, R.E., Simon, C. and Perrey, S.W. (2000) The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Mol Biol Evol* **17**: 530-539.
- Higgs, P.G. (1998) Compensatory neutral mutation and the evolution of RNA. *Genetica* **102**: 91-101.
- Higgs, P.G. (2000) RNA secondary structure: Physical and computational aspects. *Q Rev Biophys* **30**: 199-253.
- Hillis, D.M. and Dixon, M.T. (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. *Q Rev Biol* **66**: 411-453.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh Chem* **125**: 167-188.
- Horn, G.H. (1889) A synopsis of the Halticini of boreal America. *Trans Amer Entomol Soc* **16**: 163-320.
- Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M. and Higgs, P.G. (2003) RNA-based phylogenetic methods: Application to mammalian mitochondrial RNA sequences. *Mol Phylogenet Evol* **28**: 241-252.
- Huelsenbeck, J.P. (1997) Is the Felsenstein zone a fly trap? *Syst Biol* **46**: 69-74.
- Huelsenbeck, J.P., Larget, B., Miller, R.E. and Ronquist, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol* **51**: 673-688.

- Hwang, S.K. and Kim, J.G. (2000) Secondary structure and phylogenetic implications of nuclear large subunit ribosomal RNA in the ectomycorrhizal fungus *Tricholoma matsutake*. *Curr Microbiol* **40**: 250-256.
- Hwang, U.I., Kim, W., Tautz, D. and Friedrich, M. (1998) Molecular phylogenetics at the Felsenstein zone: Approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences. *Mol Phylogenet Evol* **9**: 470-480.
- Jackson, S.A., Cannone, J.J., Lee, J.C., Gutell, R.R. and Woodson, S.A. (2002) Distribution of rRNA introns in the three-dimensional structure of the ribosome. *J Mol Biol* **323**: 35-52.
- Jacoby, M. (1908) Coleoptera: Chrysomelidae. In *Fauna of British India Including Ceylon and Burma* Vol. 1. (Bingham, C.T., ed.), pp. 534. Taylor and Francis, London.
- Johnston, J.S., Ross, L.D., Bean, L., Hughes, D.P. and Kathirithamby, J. (2004) Tiny genomes and endoreduplication in Strepsiptera. *Insect Mol Biol* **13**: 581-585.
- Jolivet, P. (1977) Selection trophique chez les Eupoda (Coleoptera, Chrysomelidae). *Bulletin Societe Linneenne de Lyon* **46**: 321-336.
- Jolivet, P. (1991) Selection trophique chez les Alticinae (Col. Chrysomelidae). *Bulletin Societe Linneenne de Lyon* **60**: 26-40, 53-72.
- Jolivet, P., and Hawkeswood, T. (1995) *Host-Plants of Chrysomelidae of the World: An Essay About Relationships Between the Leaf Beetles and Their Food Plants*, Backhuys Publishers, Leiden, the Netherlands. 281 pp.
- Jow, H., Hudelot, C., Rattay, M. and Higgs, P.G. (2002) Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol Biol Evol* **19**: 1591-1601.
- Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* (Munro, N.H., ed.), pp. 21-132. Academic Press, New York.
- Kathirithamby, J. (1989) Review of the order Strepsiptera. *Syst Entomol* **14**: 41-92.
- Kathirithamby, J. (1991a) *Stichotrema robertsoni* spec. n. (Strepsiptera: Myrmecolacidae): The first report of stylopization in minor workers of an ant (*Pheidole* sp.: Hymenoptera: Formicidae). *J Entomol Soc South Afr* **54**: 9-15.

- Kathirithamby, J. (1991b) Strepsiptera. In *The Insects of Australia: A Textbook for Students and Research Workers*, Vol. 1 (Naumann, I.D., Carne, P.B., Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J. and Littlejohn, M.J., eds), pp. 684-695. CSIRO, Melbourne University Press, Melbourne.
- Kathirithamby, J. and Hamilton, W.D. (1992) More covert sex: The elusive females of Myrmecolacidae. *Trends Ecol Evol* **7**: 349-351.
- Kathirithamby, J., Ross, L.D. and Johnston, J.S. (2003) Masquerading as self?: Endoparasitic Strepsiptera (Insecta) enclose themselves in host-derived epidermal bag. *Proc Nat Acad Sci USA* **100**: 7655-7659.
- Kim, S.J., Kjer, K.M. and Duckett, C.N. (2003) Comparison between molecular and morphological-based phylogenies of galerucine/alticine leaf beetles (Coleoptera, Chrysomelidae: Galerucinae). *Insect Syst Evol* **34**: 53-64.
- Kimura, M. (1983) The neutral theory of molecular evolution. Cambridge University Press, New York.
- Kimura, M. (1985) The role of compensatory neutral mutations in molecular evolution. *J Genet* **64**: 7-19.
- Kimura, M. (1986) DNA and the neutral theory. *Phil Trans Roy Soc London B* **312**: 343-354.
- Kinzelbach, R.K. (1971) *Morphologische Befunde an Fächerflüglern und ihre phylogenetische Bedeutung (Insecta: Strepsiptera)*. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, Germany. 256pp.
- Kinzelbach, R. (1990) The systematic position of Strepsiptera (Insecta). *Am Entomol* **36**: 292-303.
- Kjer KM (2004) Aligned 18S and insect phylogeny. *Syst Biol* **53**: 506-514.
- Kjer, K.M. (1995) Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. *Mol Phylogenet Evol* **4**: 314-330.
- Kjer, K.M. (1997) An alignment template for amphibian 12S rRNA, domain III: conserved primary and secondary structural motifs. *J Herpetology* **31**: 599-604.
- Kjer, K.M., Baldrige, G.D. and Fallon, A.M. (1994) Mosquito large subunit ribosomal RNA: Simultaneous alignment of primary and secondary structure. *Biochim Biophys Acta* **1217**: 147-155.

- Kjer, K.M., Blahnik, R.J. and Holzenthal, R.W. (2001) Phylogeny of Trichoptera (Caddisflies): Characterization of signal and noise within multiple datasets. *Syst Biol* **50**: 781-816.
- Konstantinov, A.S. and Vandenberg, N.J. (1996) Handbook of Palearctic flea beetles (Coleoptera: Chrysomelidae: Alticinae). *Contrib Entomol Internatl* **1**: 237-439.
- Kraus, F., Jarecki, L., Miyamoto, M., Tanhauser, S. and Laipis, P. (1992) Mispairing and compensational changes during the evolution of mitochondrial ribosomal RNA. *Mol Biol Evol* **9**: 770-774.
- Kretzer, A., Li, Y., Szaro, T. and Bruns, T.D. (1996) Internal transcribed spacer sequences from 38 recognized species of *Suillus sensu lato*: Phylogenetic and taxonomic implications. *Mycologia* **88**: 776-785.
- Kuzoff, R.K., Swere, J.A., Soltis, D.E., Soltis, P.S. and Zimmer, E.A. (1998) The phylogenetic potential of entire 26S rDNA sequences in plants. *Mol Biol Evol* **15**: 251-263.
- Kwon, O.Y., Ogino, K. and Ishikawa, H. (1991) The longest 18S ribosomal RNA ever known. *Eur J Biochem* **202**: 827-833.
- Laboisière, V. (1921) Etude des Galerucini de la collection du Musée Congo belge. *Rev Zool Africaine* **9**: 33-86.
- Larget, B. and Simon, D. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol Biol Evol* **16**: 750-759.
- Larsen, N. (1992) Higher order interactions in 23S rRNA. *Proc Natl Acad Sci USA* **89**: 5044-5048.
- Larsen, N., Olsen, G.J., Maidak, B.L., McCaughey, M.J., Overbeek, R., Macke, T.J., Marsh, T.L. and Woese, C.R. (1993) The ribosomal database project. *Nucleic Acids Res* **21**: 3021-3023.
- Latreille, P.A. (1802) *Histoire naturelle, générale et particulière des Crustacés et des Insectes. Ouvrage faisant suite à l'Histoire Naturelle générale et particulière, composée par Leclerc de Buffon, et rédigé par C. S. Sonnini, membre de plusieurs Sociétés savantes. Familles naturelles des Genres*. Vol. 3. 467 pp. Dufart, Paris.
- Lawrence, J.F. and Britton, E.B. (1994) *Australian Beetles*. 192 pp. Melbourne University Press, Melbourne.

- Lee, M.S.Y. (2001) Unalignable sequences and molecular evolution. *Trends Ecol Evol* **16**: 681-685.
- Leng, C.W. (1920) *Catalogue of the Coleoptera of America, North of Mexico*. Sherman, Mount Vernon, NY. 470 pp.
- Levinson, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**: 203-221.
- Li, W. H. and X. Gu. (1996) Estimating evolutionary distances between DNA sequences. *Methods Enzymol* **266**:449-459.
- Lingafelter, S.W. and Konstantinov, A.S. (2000) The monophyly and relative rank of alticine and galerucine leaf beetles: A cladistic analysis using morphological characters (Coleoptera: Chrysomelidae). *Entomol Scand* **30**: 397-416.
- Linhart, H. and Zucchini, W. (1986) *Model Selection*. John Wiley & Sons, New York.
- Lopatin, I.K. (1984) *Leaf Beetles (Chrysomelidae) of Central Asia and Kazakhstan*. Oxonian Press, New Dehli. 416 pp.
- Lunt, D.H., Zhang, D.X., Szymura, J.M. and Hewitt, G.M. (1996) The insect cytochrome oxidase I gene: Evolutionary patterns and conserved primers for phylogenetic studies. *Insect Mol Biol* **5**: 153-65.
- Lutzoni, F., Wagner, P. and Reeb, V. (2000) Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses using unequivocal coding and optimal character-state weighting. *Syst Biol* **49**: 628-651.
- Lydeard, C., Holznagel, W.E., Schnare, M.N. and Gutell, R.R. (2000) Phylogenetic analysis of molluscan mitochondrial LSU rDNA sequences and secondary structures. *Mol Phylogenet Evol* **15**: 83-102.
- Maddison, D.R. and Maddison, W.P. (2000) *MacClade 4: Analysis of Phylogeny and Character Evolution*. Version 4.0. Sinauer Associates, Sunderland, MA.
- Mallatt, J., Garey, J.R. and Shultz, J.W. (2004) Ecdysozoan phylogeny and Bayesian inference: First use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol Phylogenet Evol* **31**: 178-191.
- Manos, P.S., (1997) Systematics of *Nothofagus* (Nothofagaceae) based on rDNA spacer sequences (ITS): Taxonomic congruence with morphology and plastid sequences. *Am J Bot* **84**: 1137-1155.

- Manuel, M., Borchellini, C., Alivon, E., Le Parco, Y., Vacelet, J. and Boury-Esnault, N. (2003) Phylogeny and evolution of calcareous sponges: Monophyly of Calcinea and Calcaronea, high level of morphological homoplasy and the primitive nature of axial symmetry. *Syst Biol* **52**: 311-333.
- Marshall, C.R. (1992) Substitution biases, weighted parsimony and amniote phylogeny as inferred from 18S-ribosomal-RNA sequences. *Mol Biol Evol* **9**: 370-377.
- Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911-940.
- Maulik, S. (1926) *The Fauna of British India, Including Ceylon and Burma. Coleoptera, Chrysomelidae (Chrysomelinae and Halticinae)*. 442 pp. Taylor and Francis, London.
- Medvedev, L.N. (1982) *Leaf Beetles of the Mongolian People's Republic: A Key for Determination*. 303 pp. Nauka Publishers, Moscow.
- Metcalf, R.L. (1994) Chemical ecology of Diabroticites. In *Novel Aspects of the Biology of the Chrysomelidae* (Jolivet, P.H., Cox, M.L. and Petitpierre, E., eds.), pp. 153-169. Kluwer Academic Publ., Boston.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of states calculations for fast computing machines. *J Chem Phys* **21**: 1087-1091.
- Michot, B., Hassouna, N. and Bachellerie, J-P. (1984) Secondary structure of mouse 28S rRNA and general model for the folding of the large rRNA in eukaryotes *Nucleic Acids Res* **12**: 4259-4279.
- Michot, B. and Bachellerie, J-P. (1987) Comparisons of large subunit rRNAs reveal some eukaryote-specific elements of secondary structure. *Biochimie* **69**: 11-23.
- Miller, R.E., McDonald, A. and Manos, P.S. (2004) Systematics of *Ipomoea* subgenus *Quamoclit* (Convolvulaceae) based on its sequence data and a Bayesian phylogenetic analysis. *Am J Bot* **91**: 1208-1218.
- Mindell, D.P. and Honeycutt, R.L. (1990) Ribosomal RNA in vertebrates: Evolution and phylogenetic implications. *Annu Rev Ecol Syst* **21**: 541-566.
- Misof, B. and Fleck, G. (2003) Comparative analysis of mt LSU rRNA secondary structures of odonates: Structural variability and phylogenetic signal. *Insect Mol Biol* **12**: 535-547.

- Mitchison, G.J. (1999) A probabilistic treatment of phylogeny and sequence alignment. *J Mol Evol* **49**: 11-22.
- Mohr, K.-H. (1966) 88. Familie: Chrysomelidae. In *Käfer, Mitteleuropas*, Vol. 9 (Freude, H, Harde, K.W., Lohse, G.A., eds.), pp. 95-280. Goecke & Evers, Krefeld, Germany.
- Morin, L. (2000) Long branch attraction effects and the status of "basal eukaryotes": Phylogeny and structural analysis of the ribosomal RNA gene cluster of the free-living diplomonad *Trepomonas agilis*. *J Eukaryot Microbiol* **47**: 167-177.
- Morrison, D.A. and Ellis, J.T. (1997) Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. *Mol Biol Evol* **14**: 428-441.
- Moulton, V., Zuker, M., Steel, M., Pointon, R. and Penny, D. (2000) Metrics on RNA secondary structures. *J Comput Biol* **7**: 277-292.
- Mugridge, N.B., Morrison, D.A., Johnson, A.M., Luton, K., Dubey, J., Votypka, J. and Tenter, A.M. (1999) Phylogenetic relationships of the genus *Frenkelia*: A review of its history and new knowledge gained from comparison of large subunit ribosomal ribonucleic acid gene sequences. *Int J Parasitol* **29**: 957-972.
- Munroe, D.D. and Smith, R.F. (1980) A revision of the systematics of *Acalymma sensu stricto* barber (Coleoptera: Chrysomelidae) from North America including Mexico. *Mem Ent Soc Canada* **112**: 1-92.
- Muse, S. (1995) Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics* **139**: 1429-1439.
- Musters, W., Venema, J., van der Linden, G., van Heerikhuizen, H., Klootwijk, J. and Planta, R.J. (1989) A system for the analysis of yeast ribosomal DNA mutations. *Mol Cell Biol* **9**: 551-559.
- Musters, W., Goncalves, P.M., Boon, K., Gaué, H.A., van Heerikhuizen, H. and Planta, R.J. (1991) The conserved GTPase center and variable region V9 from *Saccharomyces cerevisiae* 26S rRNA can be replaced by their equivalents from other prokaryotes or eukaryotes without detectable loss of ribosomal function. *Proc Natl Acad Sci USA* **88**: 1469-1473.
- Nagl, W. (1978) *Endopolyploidy and Polyteny in Differentiation and Evolution: Towards an Understanding of Quantitative and Qualitative Variation of Nuclear DNA in Ontogeny and Phylogeny*. Elsevier/North-Holland Biomed, New York.

- Nagylaki, T. (1988) Gene conversion, linkage and the evolution of multigene families. *Genetics* **120**: 291-301.
- Nedbal, M.A., Allard, M.W. and Honeycutt, R.L. (1994) Molecular systematics of hystricognath rodents: Evidence from the mitochondrial 12S rRNA gene. *Mol Phylogenet Evol* **3**: 206-220.
- Newman, E. (1835) Attempted division of British insects into natural orders. *Entomol Mag* **2**: 379-431.
- Nielsen, J.K. (1988) Crucifer-feeding Chrysomelidae: Mechanisms of host plant finding and acceptance. In *Biology of the Chrysomelidae* (Jolivet, P., Petitpierre, E. and Hsiao, T.S., eds.), pp. 25-40. Kluwer Acad. Publ., Dordrecht.
- Noller, H.F. (1991) Ribosomal RNA and translation. *Ann Rev Biochem* **60**:191-227.
- Notredame, C., O'Brien, E.A. and Higgins, D.G. (1997) RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res* **25**: 4570-4580.
- Ogloblin, A.A. (1939) The Strepsiptera parasites of ants. *Int Congr Entomol Berlin (1938)* **2**: 1277-1284.
- Ogloblin, D.A. (1936) *Leaf beetles, Galerucinae. Fauna of the USSR*. Insecta, Coleoptera Vol. 26. 455 pp. Zoological Institute of the Academy of Sciences of the USSR, Moscow.
- Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246**: 96-98.
- Ohta, T. (1980) *Evolution and Variation of Multigene Families*. Springer-Verlag, Berlin.
- Ohta, T. (1982) Allelic and nonallelic homology of a supergene family. *Proc Natl Acad Sci USA* **79**: 3251-3254.
- Ouvrard, D., Campbell, B.C., Bourgoin, T. and Chan, K..L. (2000) 18S rRNA secondary structure and phylogenetic position of Peloridiidae (Insecta, Hemiptera). *Mol Phylogenet Evol* **16**: 403-417.
- Page, R.D.M., Crulckshank, R. and Johnson, K.P. (2002) Louse (Insecta: Phthiraptera) mitochondrial 12S rRNA secondary structure is highly variable. *Insect Mol Biol* **11**: 361-369.

- Petersen, G., Seberg, O., Aagesen, L. and Frederiksen, S. (2004) An empirical test of the treatment of indels during optimization alignment based on the phylogeny of the genus *Secale* (Poaceae). *Mol Phylogenet Evol* **30**: 733-742.
- Petrov, D.A. (2001) Evolution of genome size: New approaches to an old problem. *Trends Genet* **17**: 23-28.
- Petrov, D.A. (2002) Mutational equilibrium model of genome size evolution. *Theor Pop Biol* **61**: 533-546.
- Petrov, D.A., Lozovskaya, E.R. and Hartl, D.L. (1996) High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**: 346-349.
- Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L. and Shaw, K.L. (2000) Evidence for DNA loss as a determinant of genome size. *Science* **287**: 1060-1062.
- Posada, D. and Crandall, K. (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14**: 817-818.
- Posada, D. and Crandall, K.A. (2001) Selecting the best-fit model of nucleotide substitution. *Syst Biol* **50**: 580-601.
- Prokopowich, C.D., Gregory, T.R. and Crease, T.J. (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46**: 48-50.
- Rambaut, A. and Drummond, A.J. (2004) Tracer ver 1.1, Available from <http://evolve.zoo.ox.ac.uk/>.
- Rannala, B. (2002) Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst Biol* **51**: 754-760.
- Redtenbacher, L. (1874) *Fauna Austriaca. Die Käfer. Dritte, ganzlich umgearbeitete und bedeutend vermehrte Auflage.* 564 pp. Erster Band, Verlag von Carl Gerold's Sohn, Vienna.
- Reid, C.A.M. (1995a) A cladistic analysis of subfamilial relationships in the Chrysomelidae *sensu lato* (Chrysomeloidea). In *Biology, Phylogeny and Classification of Coleoptera: Papers celebrating the 80th birthday of Roy A. Crowson* (Pakaluk, J. and Slipinski, S.A., eds.), pp. 559-632. Muzeum i Instytut Zoologii PAN, Warszawa.
- Reid, C.A.M. (1995b) Errors noted, changes needed. *Chrysomela*, **30**, 4.

- Reid, C.A.M. (2000) Spilopyrinae Chapuis: A new subfamily in the Chrysomelidae and its systematic placement (Coleoptera). *Invert Tax* **14**: 837-862.
- Riley, E.G., Clark, S.M., Flowers, R.W. & Gilbert, A.J. (2002) Chrysomelidae Latreille 1802. In *American Beetles*, Vol 2., *Polyphaga: Scarabaeoidea through Curculionoidea* (Arnett, R.H., Thomas, M.C., Skelley, P.E. and Frank, J.H., eds.), pp. 617-691. CRC Press, Boca Raton, FL.
- Rimoldi, O.J., Raghu, B., Mag, M.K. and Eliceiri, G.L. (1993) Three new small nucleolar RNAs that are psoralen cross-linked *in vivo* to unique regions of pre-rRNA. *Mol Cell Biol* **13**: 4382-4390.
- Ronquist, F. and Huelsenbeck, J.P. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572-1574.
- Rousset, F., Pelandakis, M. and Solignac, M. (1991) Evolution of compensatory substitutions through GU intermediate state in *Drosophila* rRNA. *Proc Natl Acad Sci USA* **88**: 10032-10036.
- Samuelson, G.A. (1994) Pollen consumption and digestion by leaf beetles. In *Novel Aspects of the Biology of the Chrysomelidae* (Jolivet, P.H., Cox, M.L. and Petitpierre, E., eds.), pp. 179-183. Kluwer Academic Publ., Boston.
- Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM J Appl Math* **28**: 35-42.
- Sankoff, D., Morel, C. and Cedergren, R.J. (1973) Evolution of 5S RNA and the non-randomness of base replacement. *Nature New Biol* **245**: 232-234.
- Sankoff, D. and Cedergren, R.J. (1983) Simultaneous comparison of three or more sequences related by a tree. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* (Sankoff, D. and Kruskal, J.B., eds.), pp. 253-263. Addison-Wesley, Reading, MA.
- Savill, N., Hoyle, D. and Higgs, P. (2001). RNA sequence evolution with secondary structure constraints: Comparison of substitution rate models using maximum likelihood methods. *Genetics* **157**: 399-411.
- Scherer, G. (1969) Die Alticinae des indischen Subkontinentes (Coleoptera - Chrysomelidae) *Pacific Insects Monograph* no. **22**, 251 pp.
- Schlutzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F. and Yonath, A. (2000) Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102**: 615-623.

- Schnare, M.N., Damberger, S.H., Gray, M.W. and Gutell, R.R. (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J Mol Biol* **256**: 701-719.
- Schöniger, M. and von Haeseler, A. (1994) A stochastic model for the evolution of autocorrelated DNA sequences. *Mol Phylogenet Evol* **3**: 240-247.
- Schultes, E.A., Hraber, P.T. and LaBean, T.H. (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J Mol Evol* **49**: 76-83.
- Seeno, T.N. & Wilcox, J.A. (1982) Leaf beetle genera (Coleoptera: Chrysomelidae). *Entomography* **1**: 1-222.
- Shapiro, B.A. and Zhang, K. (1990) Comparing multiple RNA secondary structures using tree comparisons. *CABIOS* **6**: 309-318.
- Silfverberg, H. (1990) The nomenclaturally correct names of some family-groups in Coleoptera. *Entomologica Fennica*, **1**: 119-121.
- Smith, A.B. (1989) RNA sequence data in phylogenetic reconstruction: Testing the limits of its resolution. *Cladistics* **5**: 321-344.
- Sorenson, M.D., Oneal, E., Garcia-Moreno, J. and Mindell, D.P. (2003) More taxa, more characters: The Hoatzin problem is still unresolved. *Mol Biol Evol* **20**: 1484-1499.
- Spahn, C.M., Beckmann, R., Eswar, N., Penczek, P.A., Sali, A., Blobel, G. and Frank, J. (2001) Structure of the 80S ribosome from *Saccharomyces cerevisiae*-tRNA-ribosome and subunit-subunit interactions. *Cell* **107**: 373-386.
- Springer, M.S., Hollar, L.J. and Burk, A. (1995) Compensatory substitutions and the evolution of the mitochondrial 12S rRNA gene in mammals. *Mol Biol Evol* **12**: 1138-1150.
- Springer, M.S. and Douzery, E. (1996) Secondary structure and patterns of evolution among mammalian mitochondrial 12S rRNA molecules. *J Mol Evol* **43**: 357-373.
- Stephan, W. (1996) The rate of compensatory evolution. *Genetics* **144**: 419-426.
- Stephens, J.F. (1839) *A manual of British Coleoptera, or beetles; containing a brief description of all the species of beetles hitherto ascertained to inhabit Great Britain and Ireland; together with a notice of their chief localities, times and places of appearances, etc.* 443 pp. Longman, Orme, Brown, Green, and Longmans, London.

- Strand, M.R. (1986) The physiological interactions of parasitoids with their hosts and their influence on reproductive strategies. In *Insect Parasitoids* (Waage, J. and Greathead, D., eds), 13th Symposium of the Royal Entomological Society of London, 18-19 September 1985, pp. 97-136.
- Strand, M.R. and Peach, L.L. (1995) Immunological basis for compatibility in parasitoid-host relationships. *Ann Rev Entomol* **40**: 31-56.
- Suzuki, K. and Furth, D.F. (1992) What is classification? A case study in insects systematics: Potential confusion before order. *Zool Sci* **9**: 1113-1126.
- Sweeney, R. and Yao, M.-C. (1989) Identifying functional regions of rRNA by insertion mutagenesis and complete gene replacement in *Tetrahymena thermophila*. *EMBO J* **8**: 933-938.
- Sweeney, R., Chen, L. and Yao, M.-C. (1994) An rRNA variable region has an evolutionary conserved essential role despite sequence divergence. *Mol Cell Biol* **14**: 4203-4215.
- Swofford, D.L. (1993) *PAUP 3: Phylogenetic Analysis Using Parsimony. User Manual*. Univ. of Illinois, Urbana.
- Swofford, D.L. (1999) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4. Sinauer Associates, Sunderland, MA.
- Swofford, D.L. (2001) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4.0b10. Sinauer Associates, Sunderland, MA.
- Tallamy, D.W., Frazier, J.L. and Mullin, C.A. (1999). An alternate route to insect pharmacophagy: The loose receptor hypothesis. *J Chem Ecol* **25**: 1987-1997.
- Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512-526.
- Tautz, D.J., Hancock, J.M., Webb, D.A., Tautz, C. and Dover, G.A. (1988) Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Mol Biol Evol* **5**: 366-376.
- Tillier, E.R.M. (1994) Maximum likelihood with multiparameter models of substitution. *J Mol Evol* **39**: 409-417.

- Tillier, E.R.M. and Collins, R.A. (1998) High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148**: 1993-2002.
- Titus, T.A. and Frost, D.R. (1996) Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). *Mol Phylogenet Evol* **6**: 49-62.
- Uchida, H., Kitae, K., Tomizawa, K.I. and Yokota, A. (1998) Comparison of the nucleotide sequence and secondary structure of the 5.8S ribosomal RNA gene of *Chlamydomonas tetragama* with those of green algae. *DNA Seq* **8**: 403-408.
- Van de Peer, Y., Neeps, J.M., DeRijk, P. and De Wachter, R. (1993) Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: Calibration of the molecular clock. *J Mol Evol* **37**: 221-232.
- Van de Peer, Y., Robbrecht, E., De Hoog, S., Caers, A., De Rijk, P. and De Wachter, R. (1999) Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Res* **27**: 179-183.
- Vawter, L. and Brown, W.M. (1993) Rates and patterns of base change in the small subunit ribosomal RNA gene. *Genetics* **134**: 597-608.
- Veldman, G.M., Klootwijk, J., De Regt, V.C.F.H., Planta, R.J., Branlant, C., Krol, A. and Ebel, J-P. (1981) The primary and secondary structure of yeast 26S rRNA. *Nucleic Acids Res* **9**: 6935-6952.
- Waddell, P.J. and Steel, M.A. (1997) General time-reversible distances with unequal rates across sites. *Mol Phylogenet Evol* **8**: 398-414.
- Walter, G.H. (1983) 'Divergent male ontogenies' in Aphelinidae (Hymenoptera: Chalcidoidea): A simplified classification and a suggested evolutionary sequence. *Biol J Linn Soc* **19**: 63-82.
- Ware V.C., Tague, B.W., Clark, C.G., Gourse, R.L., Brand, R.C. and Gerbi, S.A. (1983) Sequence analysis of 28S ribosomal DNA from the amphibian *Xenopus laevis*. *Nucleic Acids Res* **11**: 7795-7817.
- Weise, J. (1923) Chrysomeliden und Coccinelliden aus Queensland. Results of Dr. E. Mjöberg's Swedish scientific expedition to Australia 1910-1913. *Archiv für Zoologie* **15**: 1-150.
- Weise, J. (1924) Chrysomelidae: 13 Galerucinae, Vol. 25 (Pars 78). Pp. 1-225 in *Coleopterum Catalogus* (Schenkling, S., ed.). W. Junk, Berlin.

- Wheeler, W.C. (1999) Fixed character states and the optimization of molecular sequence data. *Cladistics* **15**: 379-385.
- Wheeler, W.C. and Honeycutt, R.L. (1988) Paired sequence difference in ribosomal RNAs: Evolutionary and phylogenetic implications. *Mol Biol Evol* **5**: 90-96.
- Wheeler, W.C. and Gladstein, D.L. (1994) MALIGN, version 1.93. American Museum of Natural History, New York.
- Whiting, M.F., Carpenter, J.C., Wheeler, Q.D. and Wheeler, W.C. (1997) The Strepsiptera problem: Phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology. *Syst Biol* **46**: 1-68.
- Wiegmann, B.M., Yeates, D.K., Thorne, J.L. and Kishino, H. (2003) Time flies, a new molecular time-scale for brachyceran fly evolution without a clock. *Syst Biol* **52**: 745-756.
- Wilcox, J.A. (1965) A synopsis of North American Galerucinae (Coleoptera: Chrysomelidae). *Bull NY St Mus Surv* **400**: 1-226
- Wilcox, J.A. (1971) *Coleopterum Catalogus Supplementa (Chrysomelidae: Galerucinae, Oidini, Galerucini, Metacyclini, Sermylini)*, Pars **78**, Fasc. 2. 2nd ed. Dr. W. Junk, Gravenhage, Netherlands.
- Wilcox, J.A. (1972a) *Coleopterum Catalogus Supplementa (Chrysomelidae: Galerucinae, Luperini: Diabroticina and Aulacophorina)*, Pars **78**, Fasc. 2. 2nd ed. Dr. W. Junk, Gravenhage, Netherlands.
- Wilcox, J.A. (1972b) *Coleopterum Catalogus Supplementa (Chrysomelidae: Galerucinae, Luperini: Luperina)*, Pars **78**, Fasc. 3. 2nd ed. Dr. W. Junk, Gravenhage, Netherlands.
- Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonnrhein, C., Hartsch, T., Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature* **407**: 327-339.
- Winker, S. and Woese, C.R. (1991) A definition of the domains Archaea, Bacteria and Eucarya in terms of small ribosomal RNA characteristics. *Syst Appl Microbiol* **14**: 305-310.
- Winkler, A. (1929) *Catalogus Coleopterum Regionis Palaearcticae*. Vol. 10, 1360 pp. A. Winkler, Vienna.

- Winnepenninckx, B., Backeljau, T. and De Wachter, R. (1995) Phylogeny of protostome worms derived from 18S rRNA sequences. *Mol Biol Evol* **12**: 641-649.
- Woese, C.R., Magrum, L.J., Gupta, R., Siegel, R.B., Stahl, D.A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J.J. and Noller, H.F. (1980) Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res* **8**: 2275-2293.
- Woese, C.R., Kandler, O. and Wheelis, M.L. (1990a) Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. *Proc Natl Acad Sci USA* **87**: 4576-4579.
- Woese, C.R., Winker, S. and Gutell, R.R. (1990b) Architecture of Ribosomal RNA: Constraints on the sequence of Tetra-loops. *Proc Natl Acad Sci (USA)* **87**: 8467-8471.
- Wool, I.G. (1986) Studies of the structure of eukaryotic (mammalian) ribosomes. In *Structure, Function and Genetics of Ribosomes* (Hardesty, J. and Kramer, G., eds.), pp. 391-411. Springer-Verlag, New York.
- Wool, I. G., Y. Endo, Y.-L. Chan and A. Gluck. (1990). Structure, function and evolution of mammalian ribosomes. In *The Ribosome: Structure, Function and Evolution* (Hill, W.E., Dahlbert, A., Garrett, R.A., Moore, P.B., Schlessinger, D. and Warner, J.R., eds.), pp. 203-214. American Society for Microbiology, Washington, D.C.
- Wuyts, J., De Rijk, P., Van de Peer, Y., Pison, G., Rousseeuw, P. and De Wachter, R. (2000) Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Res* **28**: 4698-4708.
- Wuyts, J., Van de Peer, Y. and De Wachter, R. (2001) Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Res* **29**: 5017-5028.
- Xia, X. (2000) Phylogenetic relationship among horseshoe crab species: The effect of substitution models on phylogenetic analyses. *Syst Biol* **49**: 87-100.
- Xia, X., Xie, Z. and Kjer, K.M. (2003) 18S ribosomal RNA and tetrapod phylogeny. *Syst Biol* **52**: 283-295.
- Yao, M.-C., Kimmel, A.R. and Gorovsky, M.A. (1974) A small number of cistrons for ribosomal RNA in the germinal nucleus of a eukaryote, *Tetrahymena pyriformis*. *Proc Nat Acad Sci USA* **71**: 3082-3086.

- Yeates, D.K. and Wiegmann, B.M. (1999) Congruence and controversy: Toward a higher-level phylogeny of the Diptera. *Annu Rev Entomol* **44**: 397-428.
- Yusupov, M.M., Yusupova, G.Z., Baucom, A., Lieberman, K., Earnest, T.N., Cate, J.H. and Noller, H.F. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**: 883-896.
- Zuker, M., Mathews, D.H. and Turner, D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In *RNA Biochemistry and Biotechnology, NATO ASI Series* (Barciszewski, J. and Clark, B.F.C., eds), pp. 11-43. Kluwer Academic Publ. Boston.

VITA

Joseph James Gillespie

**Department of Entomology
Texas A&M University
College Station, TX 77843-2475
pvittata@hotmail.com**

Education

- **2001-2005:** Texas A&M University, College Station, Texas. **Ph.D.**, Entomology. Major Advisor: Anthony I. Cognato. Dissertation: *Structure-based methods for the phylogenetic analysis of ribosomal RNA (rRNA) molecules.*
- **1998-2001:** University of Delaware, Newark, Delaware. **MS**, Entomology and Applied Ecology. Major Advisor: Douglas W. Tallamy. Thesis: *Inferring phylogenetic relationships among basal taxa of the tribe Luperini, or "rootworms" (Coleoptera: Chrysomelidae; Galerucinae), through the analysis of mitochondrial and nuclear DNA sequences.*
- **1995-1998:** Widener University, Chester, Pennsylvania. **BS**, Biology. Major Advisor: Bruce W. Grant. Independent research study: *Effect of a recent fire on arthropod diversity and abundance in the New Jersey Pine Barrens.*

Professional Experience in Biology

- **2005-2006:** Postdoctoral research assistant to Robert Wharton, Texas A&M University. Molecular phylogenetic analysis of Hymenoptera.
- **2004-2005:** Graduate Teaching Assistant, Texas A&M University. Five sections in General Entomology, 125 students. Fall 2004-Spring 2005.
- **2001-2004:** Graduate Research Assistant, Texas A&M University. DNA sequencing and molecular phylogenetic analysis of insects. Specializing in secondary-structure prediction of ribosomal RNA-genes of insects and other animals. Creation of tools for phylogenetic analysis of rRNA sequences (the jRNA project).
- **2001-2001:** Chemist, CIBA Chemicals, Newport, Delaware. HPLC, GC, analytical chemistry, laboratory maintenance.
- **1998-2001:** Graduate Research Assistant, University of Delaware. Field techniques for rearing treehoppers. DNA sequencing and molecular phylogenetic analysis of treehoppers and chrysomelid beetles. Molecular biology tutor.
- **1997-1998:** Undergraduate Research Assistant, Widener University. Field collecting techniques and identification of New Jersey Pine Barrens arthropods. Statistical calculations of arthropod diversity.